# Voir, c'est percevoir : le rôle des lèvres dans la production et la perception du /r/ anglo-anglais

Hannah King

Université de Paris

Ecole doctorale 622 : Sciences du Langage

*CLILLAC-ARP, EA 3967*

# Seeing is perceiving: The role of the lips in the production and perception of Anglo-English /r/

Hannah King

Thèse de doctorat de linguistique

Dirigée par Ioana Chitoran
Et par Emmanuel Ferragne

Présentée et soutenue publiquement le 13 novembre 2020

Devant un jury composé de :

| | | |
|---|---|---|
| Ioana Chitoran | Professeure, Université de Paris | Directrice |
| Emmanuel Ferragne | Maître de conférences, Université de Paris | Directeur |
| Sophie Herment | Professeure, Aix-Marseille Université | Présidente |
| Barbara Kühnert | Maîtresse de conférences, Université Sorbonne Nouvelle | Examinatrice |
| Rachid Ridouane | Directeur de recherche, CNRS – Université Sorbonne Nouvelle | Rapporteur |
| James Scobbie | Professor, Queen Margaret University | Rapporteur |

# Seeing is perceiving: The role of the lips in the production and perception of Anglo-English /r/

**Abstract:** Articulatory variation is well-documented in post-alveolar approximant realisations of /r/ in rhotic Englishes, which present a diverse array of tongue configurations. However, the production of /r/ remains enigmatic, especially concerning non-rhotic Englishes and the accompanying labial gesture, both of which tend to be overlooked in the literature. This thesis attempts to account for them both by considering the production and perception of /r/ in the non-rhotic variety of English spoken in England, *Anglo-English*. This variety is of particular interest because non-lingual labiodental articulations of /r/ are rapidly gaining currency, which may be due to the visual prominence of the lips, although a detailed phonetic description of this change in progress has yet to be undertaken.

Three production and perception experiments were conducted to investigate the role of the lips in Anglo-English /r/. The results indicate that the presence of labiodental /r/ has resulted in auditory ambiguity with /w/ in Anglo-English. In order to maintain a perceptual contrast between /r/ and /w/, it is argued that Anglo-English speakers use their lips to enhance the perceptual saliency of /r/ in both the auditory and visual domains. The results indicate that visual cues of the speaker's lips are more prominent than the auditory ones and that these visual cues dominate the perception of the contrast when the auditory and visual cues are mismatched. The results have theoretical implications for the nature of speech perception in general, as well as for the role of visual speech cues in diachronic sound change.

**Key words:** English; rhotics; articulation; labialisation; ultrasound tongue imaging; audio-visual speech perception; sound change

# Voir, c'est percevoir : le rôle des lèvres dans la production et la perception du /ɹ/ anglo-anglais

**Résumé :** La variabilité articulatoire dans les réalisations approximantes post-alvéolaires du /ɹ/ est bien documentée dans les variétés rhotiques de l'anglais, qui présentent une vaste palette de configurations linguales possibles. Cependant, la production du /ɹ/ reste énigmatique, notamment en ce qui concerne les variétés non-rhotiques et le geste articulatoire labial – ces derniers étant généralement négligés dans les études phonétiques. Cette thèse a pour but de prendre en compte ces deux éléments en étudiant la production ainsi que la perception du /ɹ/ dans la variété non-rhotique de l'anglais d'Angleterre, *l'anglo-anglais*. Une attention particulière mérite d'être portée à cette variété car les variantes labiodentales non-linguales commencent à s'y développer fortement. L'indice visuel important fourni par les lèvres a possiblement provoqué ce changement linguistique, dont une description phonétique détaillée n'existe toutefois pas encore.

Trois études de production et de perception ont été réalisées pour étudier le rôle des lèvres dans le /ɹ/ anglo-anglais. D'après les résultats, une réalisation labiodentale du /ɹ/ entraîne une ambiguïté auditive avec le /w/ en anglo-anglais. Afin de maintenir un contraste perceptif entre /ɹ/ et /w/, nous suggérons que les locuteurs d'anglo-anglais utilisent leurs lèvres pour augmenter la saillance perceptive du /ɹ/ dans les domaines auditif et visuel. Les résultats montrent que les indices visuels des lèvres occupent une place plus importante que les indices auditifs dans la perception du contraste entre /ɹ/ et /w/. En cas de conflit entre indices auditifs et visuels, ce sont ces derniers qui l'emportent. Ces résultats ont des implications théoriques concernant la nature de la perception de la parole en général, ainsi que le rôle des indices visuels de la parole dans les changements phonétiques diachroniques.

**Mots clefs :** anglais ; rhotiques ; articulation ; labialisation ; échographie linguale ; perception audio-visuelle ; changement phonétique

*For Susan Grace & Grace Charlotte*

*With love & gratitude*

> *Phonological, phonetic, and dialectological accounts of* [ɪ] *which neglect the contribution of lip protrusion to its production may be incomplete and present a somewhat skewed view of the physical basis of this variant.*

Docherty and Foulkes (2001), pp. 182-183

# Acknowledgements

Just as it takes a village to raise a child, it takes a village to complete a PhD. Today is the day I can finally thank the village of people who have made it possible.

I wish to express my sincere and heartfelt thanks first and foremost to my supervisors Ioana Chitoran and Emmanuel Ferragne. Their unwavering faith and confidence in me has pushed me forwards from the very beginning when I embarked on a master's in phonetics and phonology with trepidation after having rejoiced rather prematurely in the end of my studies upon completing my undergraduate degree two years previously. Even in those early days, they allowed me free rein with a brand new ultrasound machine, which would become the source of frustration and fascination in equal measure for the years to come. Ioana and Emmanuel's open-door policy, combined with their wisdom, patience and good humour have kept me grounded and smiling throughout the uncertainty that is doctoral studies and beyond. Ioana has not only provided countless fruitful discussions, but has helped me manage the stress of combining research, teaching and personal life. She is an exceptional educator, a patient mentor and an inspirational researcher. Emmanuel has taught me to think outside the (black!) box and to embrace and relish in a challenge with determination, creativity and music. He also deserves the award for being the most reactive supervisor in the business! I owe the majority of my publications and conference presentations to his persistent faith and confidence

in me. If I get the chance to continue in academia, Ioana and Emmanuel will forever remain my inspiration.

I wish to thank Sophie Herment, Barbara Kühnert, Rachid Ridouane and Jim Scobbie for giving me the honour of accepting to be members of the jury to examine this work. I am particularly grateful to my committee members Barbara Kühnert and Eleanor Lawson whose input and support throughout this process have been invaluable.

I am indebted to the Clinical Audiology, Speech and Language Research Centre at Queen Margaret University for generously allowing me to collect production data in their facilities. I was only with them for a short time but the warm welcome I received made me feel completely at home. I particularly wish to thank Eleanor Lawson, Jim Scobbie, Alan Wrench and Steve Cowen for their support and guidance with data collection and analysis. I express my thanks to the University of York, and to Paul Foulkes in particular, for allowing me to conduct my perception experiment in their linguistics department.

A large part of what has made the last four years possible is the people with whom I shared the experience on a daily basis. I wish to thank the academic staff in the linguistics and the English departments at the Université de Paris who have guided and supported me along the way; Hiyon Yoo, Georges Boulakia, Ewan Dunbar, Harim Kwon, Sylvain Navarro and Anne Talbot to name but a few. I was lucky to share the experience with past and present PhD students Rachel, Anisia, Anqi, Qianwen, Darya, Patricia and Ismaël in the ARP lab. They kept me sane with countless tea and lunch breaks and made coming to the lab a real joy. Their contributions to group discussions have also proven invaluable.

I wish to thank all the participants who provided their precious data to make this thesis possible. I'd like to recognise the assistance I received from Annabel Smith who let me recruit her students during lesson time at Harrogate College. I'd also like to express my gratitude to my former colleagues at the Centre of English Studies in Harrogate for participating during work hours and for their support through the years.

This thesis would not have seen the light of day had it not been for the support of my wonderful friends and family. Special thanks go to my Erasmus family Vera, Cathy, Sjoerdje,

Thade and Kristóf, whose friendship has remained a constant source of pride and joy ever since our Rouen days. Here's to our next reunion and 10 year friendversary!

A mes amis Lucas, Witold, Corentin et Silvia, les moments de 'PLS' étaient moins nombreux grâce à vous ! Merci pour les soirées et les vacances, pour votre bon humeur (malgré les 'trolls' !) et pour votre amitié précieuse.

To my close friends Alice, Andrew, Ella, and Elly, you have patiently endured countless chats about lips and tongues, have been poked and prodded (sorry, Crumpet!) and have shown a true interest in what I do. You have been there for all the ups and downs with a bottle of wine in hand and a smile on your face. I couldn't ask for more supportive friends. Thank you all so very much.

To Dad and Adam, the most popular Kings in Yorkshire, you should perhaps consider a career change to recruitment! Thank you for your love and support now and always. Dad, an honourable mention must go to your CSE (grade 2) in woodwork. I'm sure it laid the foundation for my academic prowess!

To Paul, thank you for listening and for taking the time to understand what it's all about. You have celebrated every success, no matter how small, with pride.

A Isabelle, Renaud, Karine, Estelle, Arthur et Chloé, merci infiniment pour votre soutien sans faille, pour votre présence au quotidien et pour les nombreux dîners, brunchs et BBQs à mes côtes. Je suis fière de faire partie de la famille. Oui, on est tous des cousins !

To the two most important and influential women in my life, Mummy and Grace, your love, support, encouragement and positivity have been boundless, despite the physical distance separating us. I dedicate this work entirely to you both.

Finally, to Ferdinand, this achievement is as much yours as it is mine. Your support has been unparalleled and I am so thankful for everything you do. Your love and partnership has got me through and for that I am eternally grateful.

# Preface

To improve the readability of this thesis, we would like to draw the reader's attention to a couple of notes on formatting:

Definitions of important terms are provided in a Glossary, which appears just before the main manuscript. In the electronic version, glossary entries are indicated in green. Clicking on one of these terms will automatically send the reader to the corresponding glossary entry via a hyperlink. A partial Index has also been provided at the end of the thesis, which lists the main occurrences of key terms.

Bibliographic references within the manuscript are also clickable. We note that first name initials have been included in the manuscript when two authors share the same surname. Wherever possible, the Digital Object Identifier (DOI) or an internet link to cited works have been provided in the bibliography, both of which are clickable.

Hypotheses for all three experiments are numbered throughout the manuscript. A clickable hyperlink (also presented in green) has been included, which allows the reader to go back to the description associated with each numbered hypothesis.

We note that '/ɾ/' is used throughout, which we consider a phonologically and phonetically neutral label for the approximant 'r' of English. While some authors use /ɹ/, /ɾ/ was preferred because it is a simpler symbol and was deemed the most neutral of the two options. Where

phonetic transcriptions need to be distinguished, we will use [ɻ] for retroflex, [ɪ] for bunched and [ʋ] for labiodental variants.

Finally, a variety of linear mixed-effects models were used for statistical analyses. Tabularised model summaries are presented for each of the discussed models and model syntax/formula has been included directly below each table.

# Table of contents

# List of figures

Chapter 1



Chapter 2



Chapter 3

CHAPTER 5

Chapter 6

# List of tables

## Chapter 2



## Chapter 4

Chapter 5



Chapter 6

# Abbreviations

**AAA** Articulate Assistant Advanced

**CNN** Convolutional Neural Network

**CU** Curled Up

**DNN** Deep Neural Network

**EMA** Electromagnetic Articulography

**EMG** Electromyography

**FB** Front Bunched

**fps** frames per second

**FU** Front Up

**H&H Theory** 'Hyper'- and 'Hypo'-articulation Theory

**MB** Mid Bunched

**MRI** Magnetic Resonance Imaging

**SNR** Signal-to-Noise Ratio

**SSBE** Standard Southern British English

**TU** Tip Up

**UTI** Ultrasound Tongue Imaging

# Glossary

**American English** The rhotic variety of English spoken in North America.

**Anglo-English** The non-rhotic variety of English spoken in England.

**approximant** A consonant whose articulators approach each other but not to such an extent as to create turbulent airflow.

**bunched** An articulation whose primary constriction occurs at the tongue dorsum. The tongue tip is generally lowered.

**clear speech (or hyperspeech)** Speech produced with the goal of improving intelligibility in the listener.

**covert articulations** Articulations which are visibly different from one another but do not produce an audible difference. Covert articulations are therefore not perceptible or recoverable from listening to the auditory signal alone.

**endolabial** A type of close lip rounding termed by Catford, which is produced with the inner surfaces of the lips. This type of rounding is associated with back vowels such as [u] and the semi-vowel [w] and is equivalent to our label **horizontal labialisation**. Another

equivalent term is **inner rounding**, coined by Sweet. As Trask describes in his *Dictionary of Phonetics and Phonology*, **outrounding** is also an unfortunate synonym.

**exolabial** A type of lip rounding termed by Catford, which is produced with the outer surfaces of the lips. This type of rounding is associated with front vowels such as [y] and is equivalent to our label **vertical labialisation**. Another equivalent term is **outer rounding**, coined by Sweet. As Trask describes in his *Dictionary of Phonetics and Phonology*, **inrounding** is also an unfortunate synonym.

**fiducial** A fixed line used as a basis of reference and measure.

**focalisation** The convergence of neighbouring formants in the spectrum of a vowel, resulting in spectral prominence in that focalised region. Vowels which exhibit **focalisation** are known as **focal** vowels and are generally considered to be more **perceptually salient** than their non-focal counterparts (Schwartz, Abry, Boë, Ménard, & Vallée, 2005).

**horizontal labialisation** A type of **labialisation** generally associated with back vowels. The lips are pouted by drawing the lip corners together to form a small, round opening.

**hyperarticulation** A type of **clear speech** which helps the listener to retrieve and decode phonetic cues. At the segmental level, hyperarticulation may involve modifications to articulation with the goal of enhancing the phonetic contrasts between sounds.

**hypercorrection** Proposed by Ohala in his perception-oriented account of sound change, the phonetically experienced listener erroneously corrects acoustic variation from the speaker, resulting in misperception. This scenario may trigger sound change when the listener turns speaker.

**hypocorrection** Proposed by Ohala in his perception-oriented account of sound change, the listener takes the acoustic signal at face value and fails to correct for phonetic variation, resulting in misperception. This scenario may trigger sound change when the listener turns speaker.

**intrusive** /ɹ/ A type of /ɹ/-**sandhi** and an extension of **linking** /ɹ/ in which /ɹ/ is pronounced at the end of words which do not end with an etymological or orthographic /r/ (e.g., *saw it* [sɔːɪ ɪt])

**labialisation** A secondary labial articulation occurring in consonants and vowels, resulting in a reduction in the overall lip area.

**linking** /ɹ/ A type of /ɹ/-**sandhi** in which /ɹ/ is pronounced in words which end with an etymological and orthographic /r/ (e.g., *car and driver* [kɑːɪ ən ˈdɹaɪvə]).

**lip protrusion** A type of **labialisation** which may accompany both **horizontal labialisation** and **vertical labialisation**. The lips are pushed forward, extending the length of the vocal tract.

**magnetic resonance imaging (MRI)** A tool for speech production research which provides dynamic images of the vocal tract in its entirety, although constriction generally images rather poorly. Recent advances in technology at the University of Southern California have increased the spatiotemporal resolution and quality of the data, capturing videos at around 83 fps, which is a dramatic increase from the previous 23 fps obtained in their earlier MRI datasets (as discussed in Toutios et al., 2016).

**McGurk Effect** A perceptual illusion occurring in incongruous audio-visual stimuli presented in the laboratory in which the listener reports hearing neither the auditory nor the visually presented sound, but a combination of the phonetic properties of the two, e.g. auditory-/ga/ combined with visual-/ba/ is perceived as /da/.

**motor equivalence** The ability to use a variety of movements to achieve the same goal under different conditions. In speech, different vocal tract shapes may be employed to achieve the same acoustic goal. For example, the primary acoustic cue of the vowel /u/ is a low second formant, which may be produced with a narrow constriction at the lips and/or at the palate. Perkell, Matthies, Svirsky, and Jordan (1993) observed a negative correlation between the two constrictions. If the palatal constriction is too large, the

labial constriction will compensate with a narrower constriction, and vice versa. This negative correlation corresponds to a phonetic **trading relation**.

**non-rhotic** A variety of English allowing /r/ to only be pronounced directly before a vowel.

**perceptual compensation** Proposed by Ohala in his perception-oriented account of sound change, the listener factors out phonetic variation from the speaker and successfully reconstructs the speaker's intended phoneme. **Perceptual compensation** prevents sound change from occurring.

**perceptually salient** Although multiple phonetic cues may be used to distinguish one sound from another, a perceptually salient one is a cue which provides particularly important information to the listener about the identity of the sound in question. Listeners are more sensitive to salient cues than they are to less salient ones and as a result, manipulations to salient speech cues would have a substantial impact on perception in the listener, contrary to changes to less salient ones.

/r/**-sandhi** A hiatus-filling (or linking) phenomenon which is generally associated with **non-rhotic** Englishes occurring at word boundaries in connected speech. In **non-rhotic** varieties, /r/ is only pronounced when directly followed by a vowel. /r/**-sandhi** is the name given to a realisation of /r/ which is not normally pronounced in an isolated word (e.g., *car* [kɑː]), but is realised in connected speech when directly followed by a word beginning with a vowel (e.g., *car and driver* [kɑːɹ ən ˈdɹaɪvə]). A distinction is made between two sub-phenomena of /r/-sandhi: **linking** /r/ and **intrusive** /r/.

**Received Pronunciation** The accent traditionally considered the prestige standard in England.

**retroflex** An articulation whose primary constriction occurs at the tongue tip. The tongue dorsum is generally lowered.

**rhotic** A variety of English allowing /r/ to be pronounced in all syllable contexts.

**semantic segmentation** A type of image classification which involves the training of a Convolutional Neural Network (CNN) to classify each pixel in an image according to a predefined set of classes.

**singular fit** A warning message occurring in linear mixed models, which is generally indicative of overfitting of the model. It often occurs when the random effects structure is too complex to be supported by the data, probably due to a lack of data.

**sublaminal** Associated with extreme **retroflex** tongue shapes, the underside of the tongue tip forms the main palatal constriction.

**sublingual space** Generally associated with apicals, particularly alveolar, dental and **retroflex** ones, a space or cavity is formed underneath the tongue when the tongue tip is raised towards the palate.

**sulcalization (or tongue-dorsum concavity)** Associated with **bunched** tongue shapes, creates a visible concave-shaped dip in the midsagittal tongue surface.

**trading relations** When different articulatory manoeuvres reciprocally contribute to a perceptually important acoustic cue, these manoeuvres may covary in order to maintain the cue in question at a constant level. As a result, dependence on one of these manoeuvres would be accompanied by less of another, and vice versa. See **motor equivalence** for an example.

**vertical labialisation** A type of **labialisation** generally associated with front vowels. The lips come together by raising the bottom lip and closing the jaw, resulting in a small, slit-like opening.

**viseme** A set of phonemes that have identical appearance on the lips, e.g., English /p/, /b/, /m/

**visual capture** A perceptual illusion occurring in incongruous audio-visual stimuli in which the listener reports hearing the visually presented sound instead of the auditory one, e.g.,

auditory-/ba/ paired with visual /va/ is perceived as /va/. Note the difference between **visual capture** and the **McGurk Effect**.

**visual enhancement** Speech perception is generally more accurate when listeners can both hear and see the speaker as opposed to just listening to them. **Visual enhancement** is the advantage for audio-visual speech compared to auditory-only speech.

# General introduction

Consider the citation from Docherty and Foulkes (2001) which provides the epigraph of this thesis. Despite almost 20 years having passed since this statement was made, the numerous phonetic, phonological and dialectal descriptions of the English post-alveolar approximant /r/ have failed to adequately account for its secondary articulation occurring at the lips. In his *Dictionary of Phonetics and Phonology*, Trask (2004) defines a secondary articulation as 'any articulation which accompanies another (primary) articulation and which involves a less radical constriction than that primary articulation, such as labialisation or velarisation' (p. 317). But for the case of English /r/, the lips may also be considered secondary in the more literal sense of the word in that they have attracted far less attention from linguists than the primary lingual articulation, and are thus overlooked in the literature. Indeed, as Docherty and Foulkes (2001) justly observe 'if its labial component is mentioned at all, it is only *en passant*' (p. 182, emphasis original). Most phonetic accounts simply state that /r/ may involve lip rounding, particularly in word-initial position. But the phonetic implementation of this so-called lip rounding has yet to be described, which is somewhat ironic given the ease with which the lips may be viewed and measured during speech, contrary to articulations occurring inside the mouth, which require more sophisticated techniques to image and analyse.

Indeed, as well as contributing to the shape and size of the vocal tract, and thus to the

acoustics of speech, the lips are a visible articulator in face-to-face communication. It has been shown countless times that speech perception may be influenced by what we see as well as by what we hear. For example, seeing a speaker's lip movements may enhance speech comprehension in adverse listening conditions by providing a complementary source of phonetic information to the auditory stream (e.g., Sumby & Pollack, 1954). However, seeing speech may not only enhance, but in some cases may *alter* what the listener hears. The most famous and arguably most dramatic demonstration of the impact of visual speech cues on auditory speech perception is the McGurk Effect, in which conflicting auditory and visual speech cues are perceived as a fusion of the two modalities (McGurk & Macdonald, 1976). Speech perception is therefore influenced by information from multiple senses, and is thus defined as *multimodal*.

This thesis attempts to address the shortfall in the literature on English /r/ by investigating the contribution of the lips to both its production and its perception in one particular variety of English, the non-rhotic English spoken in England, which we will refer to as *Anglo-English*. Just like the treatment of the lips, Anglo-English is also underrepresented in the phonetic literature on /r/. However, the lips may be particularly important to the production and perception of prevocalic /r/ in this variety. This is because a change in progress is underway in which the post-alveolar lingual articulation for /r/ is dropped/replaced for a labiodental one. Much of the fascination with the articulation of English /r/ held by linguists the world over stems from the variation it entails, particularly in the large array of possible tongue shapes with which it may be produced. There is a (mis)conception that the lingual articulation of the post-alveolar Anglo-English /r/ is less variable than in other varieties, despite a notable absence of empirical evidence. By considering the articulation of both the lips and tongue in this variety, as well as its perception in native speakers, we will not only dispute this claim, but will show that Anglo-English /r/ warrants our attention, particularly regarding its labial articulation.

In a series of three experiments, we will show that the lips may enhance both the production and the perception of Anglo-English /r/. We find that speakers actively control the articulatory parameters available to them in order to enhance the perceptibility of /r/, including increased labiality. However, exposure to labiodental /r/ without a lingual constriction has resulted in

perceptual uncertainty in England, particularly due to the acoustic proximity of labiodental /r/ ([ʋ]) with labial-velar /w/. Listeners have to tolerate such a high degree of acoustic variation for /r/ that even canonical productions of /w/ may be reconstructed as /r/ in perception. We suggest that Anglo-English /r/ has developed a specific labial gesture in order to increase its perceptibility in both the auditory and the visual domains. Perception data reveal that visual cues are more salient than the auditory ones for the /r/-/w/ contrast in Anglo-English and that seeing the speaker's lips may even override the auditory perception of the contrast. We conclude that in cases of auditory ambiguity, listeners may look to phonetic cues from the speaker's face to better disambiguate the contrast, which may have sparked the change from a variable to a more generalised labial posture in lingual productions of Anglo-English /r/. The results presented in this thesis therefore have theoretical implications for the nature of speech perception as a multimodal entity and we conclude that visual speech cues may play a role in the shaping of phonological sound systems.

This thesis is divided into three parts. In Part I, we review the existing literature which will serve as a background. In Chapter 1, we focus our attention on audio-visual speech perception, notably the effect of seeing the speaker's lip movements on the perception of spoken utterances. We end the chapter by considering the implications of multimodal speech perception to what we know about how spoken language has evolved and how it may continue to evolve. In Chapter 2, we review the existing literature on the articulation of /r/ in both rhotic and non-rhotic varieties of English. We examine the phonetic, physiological and sociolinguistic factors which may influence tongue shape, as well as provide an overview of the acoustics of /r/. We end the chapter by considering the emergence of labiodental variants in Anglo-English. In Chapter 3, we study existing phonetic accounts of labialisation in consonants and vowels in a variety of languages. Our review of the literature will lead us to call into question the appropriateness of the term *lip rounding* in phonetic descriptions of vowels and consonants and we propose that *labialisation* is a more appropriate term. We end Part I with a presentation of the motivations for the present thesis, as well as the main research questions to arise from the literature review.

In Part II, we investigate the contribution of the lips to the production of Anglo-English /r/ in two experiments. Experiment 1, which is presented in Chapter 4, examines to what extent lip protrusion contributes to the production of /r/ by considering both hyper- and non-hyperarticulated productions of /r/. We present articulatory data from Ultrasound Tongue Imaging (UTI) and synchronised lip camera videos, as well as acoustic data. In Chapter 5, we present the results from Experiment 2, in which we compare the configuration of the lips for Anglo-English /r/ and /w/ from lip camera videos using a variety of measures including techniques from deep learning.

In Part III, we investigate the contribution of the lips to the perception of Anglo-English /r/. In the final experiment of the thesis, Experiment 3, which is presented in Chapter 6, we assess to what extent the labial gesture for /r/ is perceptually salient in Anglo-English speakers by considering the perception of /r/ and /w/ in auditory-only, visual-only, congruous audio-visual and incongruous audio-visual modalities. We end this thesis with Chapter 7 in which we present a general discussion of the results, their theoretical implications and possible future directions.

*Part* I

# Background*

*Portions of this work were published in King, H. & Ferragne, E. (2020). Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English. *Journal of Phonetics, 80*, 100978. doi:10.1016/j.wocn.2020.100978

# 1 Audio-visual speech perception

Denes and Pinson's (1993) classic 'Speech Chain' of language processing depicts the chain of events associated with the communication of a spoken message from its conceptualisation in the speaker's brain to its reception and comprehension in the listener's, as presented in Figure 1.1. The central link and the only physical element connecting the speaker to the listener within this chain is the acoustic signal generated by the speaker's vocal movements. However, Denes and Pinson's Speech Chain was recently recreated by Peelle (2019) to incorporate an additional physical component of speech: the speaker's facial movements. These facial movements are visually transmitted to the listener which, like the acoustic signal, are also decoded in the listener's brain. Indeed, in the vast majority of face-to-face interactions, the listener has access to both the auditory and the visual speech cues generated by the speaker (Gagné, Rochette, & Charest, 2002) and research has consistently shown that listeners use information from the speaker's face in these interactions (Rosenblum, 2008b). This chapter will show that the addition of visual cues from the speaker's face not only facilitates communication, but may influence the auditory perception of speech and in some cases, may even contribute to language evolution and change.

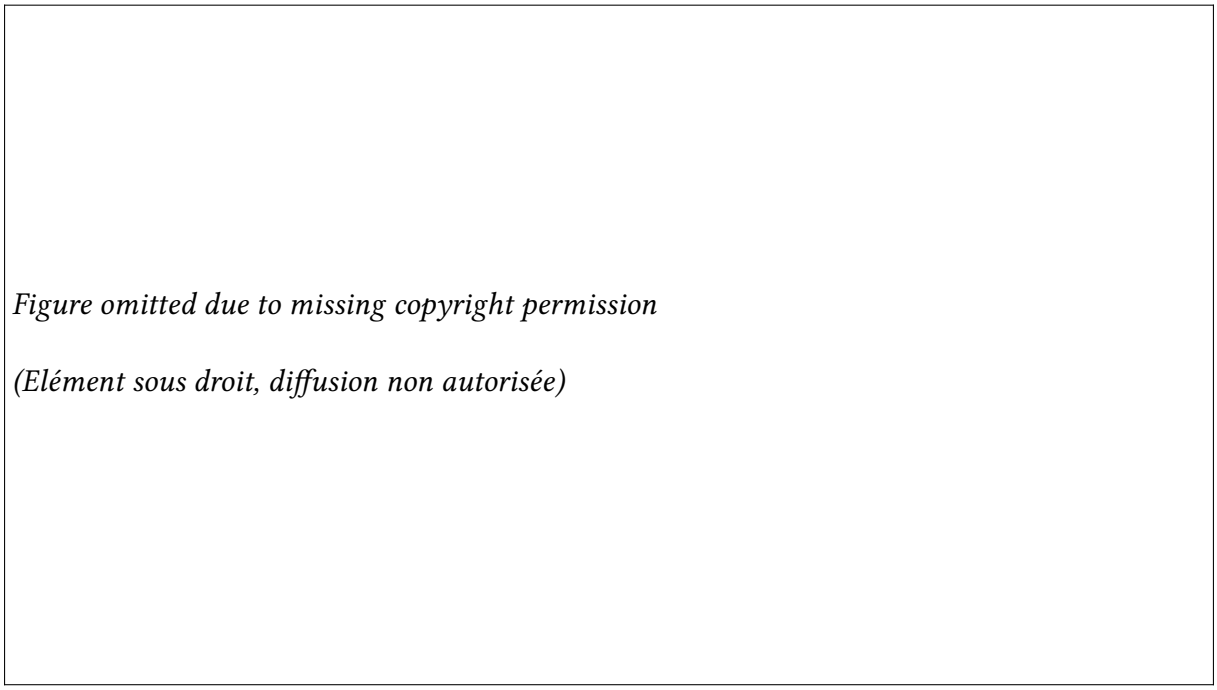



**Figure 1.1:** *Denes and Pinson's (1993) Speech Chain depicting the progression of a speech message from the brain of the speaker to the brain of the listener through the sound waves generated by the speaker's vocal movements.*

## 1.1 INFORMATION PROVIDED BY THE VISIBLE ARTICULATORS

Although non-facial movements involving the head, hands and in some respects, the entire body are used in a meaningful way in face-to-face communication, whether that be in signed or in spoken languages, we will focus on the articulatory information provided by the face, and more specifically the lips, and the role it plays in the perception of speech sounds. Indeed, most of the research on visible speech signals concentrates on the movements of the lower face, which convey the primary articulatory cues to speech events (Brooke, 1998). However, for the sake of completeness, we wish to mention the fact that certain body movements, which are not directly related to speech articulation, have been shown to convey supplementary prosodic cues to the auditory ones. For example, movements of both the head and eyebrows are used for the visual prosodic cues, or 'visual prosody' (Graf, Costatto, Strom, & Huang, 2002), involved in stress, prominence, rhythm and phrasing (e.g., Cvejic, Kim, Davis, & Gibert, 2010;

Granström, House, & Lundeberg, 1999; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Scarborough, Dmitrieva, Hall-Lew, Zhao, & Brenier, 2007).

Turning our attention to primary articulatory cues, at the most basic level, seeing the movements of the visible articulators, i.e., the lips, jaw, face, tongue tip and teeth (Badin, Tarabalka, Elisei, & Bailly, 2010), indicates to the listener that a person is speaking. This visual information is particularly useful in noisy conditions where knowing when a person is speaking enables listeners to direct their attention to the target signal. Fitzroy et al. (2018) recorded Electroencephalography (EEG) data and compared the auditory evoked potentials elicited by acoustic onsets in attended and unattended live speech in a room with multiple live speakers. Their results indicated that a visible talker is both easier to perceptually attend and harder to perceptually ignore than an unseen one. In noisy conditions, seeing that a person is speaking has also been found to aid segmentation of multiple auditory streams (Castellanos, Benedí, & Casacuberta, 1996, cited in Peelle and Sommers 2015).

However, the contribution of the visual speech cues generated by a talker in face-to-face interactions far exceeds just facilitating attention to the speaker. By presenting information about the position of a speaker's articulators, visible speech gestures may provide cues to the place of articulation of vowels and the place and manner of articulation of consonants (Summerfield, 1983, cited in Hazan et al. 2006). Visual cues of place of articulation may be particularly beneficial when the auditory conditions are degraded, e.g., due to hearing loss or environmental noise. As the acoustic cues for place of articulation are easily masked in noise, visual cues may actually be more robust than acoustic ones in some cases (Brooke, 1998). The availability of place information in the visual signal thus provides a complementary source of information to the auditory one (Peelle & Sommers, 2015) and may allow for enhanced perception of phonetic contrasts which are not very audible but are very visible, such as [m]-[n]. Contrary to the cues for place of articulation, cues for manner of articulation and voicing are not very visible but are very audible. As a result, Summerfield (1983) suggested that there is 'a fortunate complementary relationship between what is lost in noise or impairment, and what can be provided by vision' (p. 183), which allows people with hearing impairments and people

communicating in noisy conditions to supplement their perception of speech with lip reading.

However, when the acoustic information from speech is masked or removed entirely and perceivers have to rely solely on visual cues, speech perception performance is heavily reduced. For example, in cases of profound or total hearing loss, very few people are capable of understanding speech fluently by lip reading alone (Summerfield, Bruce, Cowey, Ellis, & Perrett, 1992). This is no doubt due to the ambiguous nature of the information provided by visual speech. In auditory speech, the phoneme is considered the minimal unit of contrast in the sound system of any given language. If you replace one phoneme with another, the meaning of the spoken word will change. The equivalent of the phoneme in the visual domain is the viseme (Fisher, 1968). Although its definition is somewhat disputed, Bear, Harvey, Theobald, and Lan (2014) have provided a working definition which states that a viseme is a set of phonemes that have identical appearance on the lips. Therefore, although one phoneme belongs to one viseme class, many phonemes may share the same viseme. For example, while the acoustic difference between realisations of /p/ and /b/ in English is readily perceptible due to contrasts in voice onset time, visually they are almost identical (Peelle & Sommers, 2015). Consequently, this many-to-one mapping between phonemes and visemes results in perceptual ambiguity in visual speech cues. At present, agreement has yet to be reached concerning the exact number of visemes in English, perhaps due to inter- and intra-speaker variation. Indeed, Bear et al. (2014) reviewed the phoneme-to-viseme maps for consonants presented in 15 previous studies and the number of visemes ranges from 4 to 10. Even at the most liberal estimate of 10, there are evidently far fewer consonant visemes than there are consonant phonemes in English, and the same can be said for the vowels. However, as Peelle and Sommers (2015) explained, although visual speech cues do not offer additional information compared to auditory-only speech for every phoneme, in many cases, visual cues may help disambiguate similar-sounding speech sounds.

## 1.2 Visual cues enhance auditory perception

A large body of research has shown that speech perception is more accurate when listeners can both hear and see a speaker as opposed to just listening to them. One of the first and most widely cited studies which explicitly demonstrated the utility of visual cues in the perception of speech was that of Sumby and Pollack (1954). In this study, a large cohort of participants ($n$ = 129) were asked to identify bi-syllabic words produced by a speaker seated in front of them. White noise at different intensity levels from 0 dB to -30 dB Signal-to-Noise Ratio (SNR) were presented to the subjects through a headset. Half of the subjects faced away from the speaker, while the other half watched the speaker's facial movements. In the absence of noise, subjects correctly identified nearly all of the bi-syllabic words in both the auditory-only and the audio-visual conditions with no obvious difference in performance between the two conditions. However, as the SNR decreased, i.e., as the speech signal became less audible, the visual cues played a more critical role in allowing subjects to accurately identify spoken words. Indeed, their results showed that adding the visual cue was the equivalent of improving the SNR by 15 dB. As a result, Sumby and Pollack concluded that visual speech cues contribute the most to speech intelligibility in noisy conditions.

The advantage for audio-visual speech compared to auditory-only speech, frequently known as *visual enhancement*, has since been replicated countless times. It is now widely accepted that visual speech is one of the most robust cues that people use when listening to speech in noisy environments (Lalonde & Werner, 2019). Seeing the speaker's face even from very far away (e.g., 30 m) has been shown to improve auditory speech recognition (Jordan & Sergeant, 2000). Speech perception may also be enhanced by visual cues in optimal listening conditions. Reisberg, McLean, and Goldfield (1987) demonstrated that vision enhances the perception of speech in a foreign language, speech produced by a non-native speaker and in semantically complex utterances (cited in Dohen, 2009). However, perceptual performance varies and the degree of sensitivity to visual speech cues has been linked to factors related to the perceiver's linguistic experience and development, age and sex, as well as the style of speech and the

visual salience of the cues presented. Visual cues are reported to be less beneficial to speech intelligibility in typically developing children than in adults (Desjardins, Rogers, & Werker, 1997; Lalonde & Werner, 2019; Ross et al., 2011), although older adults have been found to show no difference in their ability to perceive audio-visual speech in noise relative to younger adults (Smayda, Van Engen, Maddox, & Chandrasekaran, 2016; Sommers, Tye-Murray, & Spehar, 2005). Both children and adults with developmental disorders such as dyslexia have been shown to present deficits in their ability to gain from visual speech information relative to those without learning disorders (e.g., van Laarhoven, Keetels, Schakel, & Vroomen, 2018). Women have also been shown to be more sensitive to visual cues than men in some studies (e.g., Dancer, Krain, Thompson, Davis, & et al, 1994; Traunmüller & Öhrström, 2007; Watson, Qiu, Chamberlain, & Li, 1996), although in others, effects of speaker sex have not been reliably observed (e.g., Auer & Bernstein, 2007; Tye-Murray, Sommers, & Spehar, 2007). Shaywitz et al. (1995) considered brain activation in male and female participants during orthographic, phonological and semantic language tasks and found that their activations significantly differ. They concluded that their data provide evidence for a sex difference in the functional organisation of the brain for language, which includes phonological processing. Differences in brain activity may thus account for the reported female advantage in lip reading and visual enhancement in audio-visual speech (Desjardins & Werker, 2004).

In a way, the perception of non-native sound contrasts could be considered to be on a par with the perception of native sounds in noisy conditions, as it puts non-native perceivers at a disadvantage to native ones. Just as the benefits of visual cues vary in the perception of native speech sounds in noise, so do the results from studies assessing the benefits of visual cues in non-native speech perception. Pereira (2013) compared the sensitivity to visual cues in the perception of English vowels in Spanish learners with that of native English speakers in auditory-only, visual-only and audio-visual modalities. The results indicated that while the native speakers performed better in the audio-visual modality than the auditory-only one, no significant difference was observed between the two modalities in the Spanish learners. However, in the visual-only modality, the learners could use visual speech cues to some extent

but failed to integrate visual information to the auditory input in the audio-visual condition. In contrast, Navarra and Soto-Faraco (2007) found that while Spanish-dominant bilinguals could not distinguish between the /e/-/ɛ/ contrast in Catalan in auditory-only presentation, when presented with the accompanying visual cues, their discrimination not only improved, but did not significantly differ from that of Catalan-dominant bilinguals. Furthermore, it has been suggested that the perceiver's native language may impact sensitivity to non-native visual speech. For example, Hazan et al. (2006) found that Spanish learners show much greater sensitivity to visual cues than Japanese learners in their audio-visual perception of the non-native labial/labiodental contrast in English, although perception did improve with the presence of visual cues in both learner groups.

Research has also linked the observed variation in the benefit of visual cues to the perceptual salience of the speech cues under presentation. In a second experiment, Hazan et al. (2006) examined the perception of the /l/-/r/ contrast in learners of English and found that neither Korean nor Japanese learners showed evidence of making use of visual cues in their perception of the contrast. The authors suggested that this lack of visual enhancement is due to the fact that the /l/-/r/ contrast is not particularly visually salient. Similar results have been observed in the perception of native speech contrasts. Traunmüller and Öhrström (2007) observed a difference in the visual enhancement effect between lip rounding and mouth opening in the perception of Swedish vowels. They presented Swedish subjects with auditory, visual and audio-visual nonsense syllables in optimal listening conditions containing rounded and non-rounded vowels of different heights. They found that subjects relied more heavily on visual cues for vowel rounding than for vowel height, which they concluded may be due to the fact that lip rounding is more visually salient than mouth opening. As a result, Traunmüller and Öhrström suggested that the perception of any given feature is dominated by the modality which provides the most reliable information. In their data, contrary to contrasts involving height, the visual modality was more salient than the acoustic one for rounding, which explains the improved perceptual performance with the presence of visual cues.

Various studies have demonstrated that speakers may well be aware of the benefits of

producing visually salient cues in improving their speech intelligibility. It has been suggested that speakers make adaptations to their articulation in noisy environments as an intentional communication strategy to facilitate the transmission of the speech signal to the listener (e.g., Fitzpatrick, Kim, & Davis, 2015). Speech adaptations in noise, known as *Lombard Speech* (Lombard, 1911), may result in changes to both acoustic (e.g., Junqua, 1993) and visual speech cues (Fitzpatrick et al., 2015), and studies have shown that these changes can make speech more intelligible to listeners (e.g., Gagné et al., 2002; Van Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). With regards to articulation, clear speech has been shown to present more salient visual cues with more extreme and greater degrees of articulatory movements, including increased lip protrusion and jaw movements (Tang et al., 2015), although strategies may be speaker-specific and not all speakers make use of the visual modality to improve their speech intelligibility in noise (Garnier, Ménard, & Alexandre, 2018). It has also been observed that clear speech improves speech intelligibility to a greater extent in audio-visual than in auditory-only speech presentation (Kim, Sironic, & Davis, 2011; Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014), suggesting that the enhanced articulatory gestures made when speaking in noise may serve to make speech more visually intelligible.

Finally, it is worth pointing out that although there is an assumption that the less auditory information available to listeners, the more they will rely on visual cues, research has shown that this is not necessarily the case. Early speech perception in noise studies have indicated that visual cues benefit speech perception the most in the noisiest of conditions (e.g., Erber, 1975; Sumby & Pollack, 1954). However, maximal benefit may actually occur midway between extreme noise and no noise at all. Contrary to past studies, Ross et al. (2011) used a large word list, which participants were not exposed to prior to experimentation. Their results indicated that word recognition is considerably poorer at low SNRs than previously shown. The maximal gain from audio-visual stimulation was found to be at an SNR of around -12 dB, where performance was up to three times higher relative to auditory-only presentation. They concluded that maximum audio-visual multisensory integration occurs between the extremes where subjects have to rely mostly on lip reading (-24 dB) and where information from articulation is largely

redundant to the auditory signal (0 dB). They therefore proposed that a minimal level of auditory input is necessary before recognition can be the most effectively enhanced by visual input.

### *1.2.1 Interim summary*

To summarise, seeing a speaker's articulatory movements from the lips, jaw, tongue tip and teeth, not only allows listeners to pay better attention to the speaker, but substantially improves speech intelligibility. The benefits from visual speech cues are most notable when the accompanying auditory information is degraded, either due to hearing loss or environmental noise. However, the highest perceptual advantage from visual speech input still requires a certain degree of auditory input. Indeed, very few people are capable of understanding speech fluently from the visual signal alone. Observations from previous research thus indicate that speech intelligibility is substantially greater in audio-visual speech than in auditory-only and visual-only speech combined. However, a variety of factors have been shown to influence the extent to which visual cues may improve auditory speech perception. These factors involve, but are not limited to, inter-subject variability including age, sex and linguistic background, as well as the visual salience of the speech cues under presentation.

## 1.3 Visual cues influence auditory perception

The term *visual enhancement* implies that although the presence of visual information improves perceptual performance, auditory information remains the primary cue to speech perception (Peelle & Sommers, 2015). Indeed, up to now our review of the literature has indicated that visual cues provide somewhat redundant information in comparison to the auditory ones. We have seen, for example, that it is predominantly when some of the auditory information is missing that visual cues come into play by supplying the missing information. In this respect, seeing a speaker's articulatory movements provides complementary information and serves to augment and enhance the listener's auditory capabilities (Ross et al., 2011). However, in the following section we will show that visual inputs may actually override auditory phoneme

perception rather than enhance it, which suggests that visual cues may hold more perceptual weight than we might have initially given them credit for.

### 1.3.1 The McGurk Effect

The fact that speech perception is multimodal is arguably most dramatically demonstrated by the well-known and oft-cited McGurk Effect. McGurk and Macdonald (1976) observed that when subjects are presented with auditory recordings of /ba/ paired with video of the lips producing /ga/, they generally report perceiving neither /ba/ nor /ga/, but /da/. McGurk and Macdonald described this illusion as *perceptual fusion*[1] because phonetic properties of both the auditory and visual cues combine to form a new, fused auditory percept. Subjects perceive neither what they see nor what they hear but something in between. McGurk and Macdonald suggested that the illusory effect is particularly robust because 98% of adult and 81% of child subjects reported perceiving the fused /da/ in the incongruous auditory-/ga/ paired with visual-/ba/ condition. Furthermore, they observed that the effect persists even when subjects are aware of it. In the cases where subjects were presented with the opposite condition, i.e., incongruous auditory-/ga/ paired with visual-/ba/, subjects more often reported hearing a combination of the two sounds, /bga/, or the visual cue only, /ba/. McGurk and Macdonald concluded that their results reflect the inadequacy of auditory-based theories of speech perception as vision clearly plays a role. The McGurk Effect has since been used as the go-to experimental paradigm for studying the mechanisms underlying audio-visual speech integration (Alsius, Paré, & Munhall, 2018).

Despite the robustness of the phenomenon according to McGurk and Macdonald (1976), reported incidence rates of the illusion vary greatly across studies. The factors which influence the magnitude of the McGurk Effect are very similar to the ones which impact visual enhancement in congruous audio-visual speech. Based on their review of existing studies which use the McGurk Effect as the experimental paradigm in English speaking participants, Alsius et al. (2018) attributed variability to the following factors: the prominence of the auditory and

[1] The term *blends* also appears in the literature.

visual signals, the quality of the talker, inter-subject variability including age, sex and linguistic background, as well as factors relating to the paradigm itself, including task instructions, response structure and phonetic specificities from the audio-visual pairings. For example, the fusion effect is strongest when the auditory signal is weak. As Alsius et al. (2018) noted, it is not surprising that the McGurk Effect tends to be an illusion of visual dominance of acoustic place of articulation cues given the fact that place of articulation is one of the weakest acoustic features of speech. However, decreasing the intensity of auditory cues or masking them with noise also results in increased incidences of the McGurk illusion (Colin, Radeau, Deltenre, Demolin, & Soquet, 2002; Fixmer & Hawkins, 1998; Sekiyama, Kanno, Miura, & Sugita, 2003). Conversely, when visibility decreases, incidences of the McGurk Effect decrease, e.g., by adding noise (Fixmer & Hawkins, 1998), by reducing image resolution (A. H. Wilson, Alsius, Paré, & Munhall, 2016), or by increasing the viewing distance (Jordan & Sergeant, 2000). According to Jordan and Sergeant (2000), in incongruous audio-visual pairings, the visual signal needs to be more informative than the auditory one to have an influence on auditory perception, whereas a degraded visual signal is sufficient to improve perception of congruous audio-visual pairings.

As Alsius et al. (2018) described, variability in the magnitude of the McGurk Effect has also been related to general factors such as age, sex and linguistic background. Children are less susceptible to the effect than adults (Burnham & Dodd, 2004; McGurk & Macdonald, 1976), women report more fused percepts than men (Aloufy, Lapidot, & Myslobodsky, 1996), and some languages may be more predisposed to the McGurk Effect than others. Fewer incidences of the McGurk Effect have been reported in Asian languages, i.e., in Japanese (Sekiyama, 1994; Sekiyama & Tohkura, 1991), Mandarin (Sekiyama, 1997) and Cantonese (Burnham & Lau, 1998). Various hypotheses exist to account for the difference between Asian and non-Asian languages and the degree to which visual cues are employed in the perception of speech. Sekiyama (1997) suggested that tonal languages require more reliance on auditory cues than non-tonal ones, thus decreasing the importance of visual cues. An alternative hypothesis is that the phonemes of Japanese and Mandarin may be easier to discriminate without visual cues than those in English, making visual cues less informative (Sekiyama & Burnham, 2008). It has also been

remarked that in Asian cultures, direct viewing of the face is considered impolite, which would discourage people from these cultures from using visual cues in their perception of speech (Sekiyama, 1997). However, basing their observations on a much higher sample of speakers than in previous studies, Magnotti et al. (2015) found similar frequencies of the McGurk Effect in Chinese and American participants and showed that the main effect of culture and language accounts for only 0.3% of the variance in the data, indicating that variability may simply be due to differences in individual susceptibility to the illusion. Their study therefore highlights the necessity for large sample sizes in estimating group differences in the effect of visual cues on speech perception. As they pointed out, lack of statistical power may also account for the variability we find across studies that compare McGurk perception across different groups with regards to sex, age and clinical populations. This may well be the case in studies that have considered visual enhancement in congruous audio-visual speech too.

### *1.3.2 Visual capture*

According to Alsius et al. (2018), another reason for the variability reported in the frequency of McGurk Effect illusions may be due to confusion surrounding its exact definition. In the original study, according to McGurk and Macdonald (1976), the effect results in either a fused percept or a combination of the auditory and visual cues. In the typical stimuli used in McGurk paradigms, where auditory /ba/ is combined with visual /ga/, perception responses of /da/ or /bga/ would thus be considered possible McGurk illusions. However, in instances in which the visual component overrides the auditory one, e.g., perceiving /ba/ in the context of auditory /ga/ paired with visual /ba/, some researchers have revised McGurk and Macdonald's original definition to incorporate these visual responses as possible manifestations of the McGurk Effect, due to the fact they are visually influenced (e.g., Colin et al., 2002; Rosenblum & Saldaña, 1992; Sams, Manninen, Surakka, Helin, & Kättö, 1998, as cited in Alsius et al. 2018). However, to distinguish between illusory audio-visual responses and visual ones, other researchers have avoided using the McGurk Effect terminology and employed instead the term *visual capture*.[2]

[2]The term *visual dominance* also appears in the literature.

Visual capture, a less well-known illusion than the McGurk Effect, occurs when listeners who are perceiving incongruous audio-visual speech report hearing the visually presented sound instead of the auditory one (Mattheyses & Verhelst, 2015). This effect is arguably even more dramatic than the McGurk Effect as 'it is in visual capture that the impact of the visible articulation of speech on the resulting percept is most obvious' (Desjardins et al., 1997, p. 86). It has been remarked that for visual capture to occur, the phonetic cues in the visual signal need to be more perceptually salient than the ones in the acoustic signal (Masapollo, Polka, & Ménard, 2017). When adults aged 18-40 years were presented with incongruous auditory-/ga/ paired with visual-/ba/ and auditory-/ka/ paired with visual-/pa/, McGurk and Macdonald (1976) found higher proportions of visual responses (31% and 37%, respectively) than auditory ones (11% and 13%, respectively), indicating that visual capture occurred in some subjects in these contexts. On the other hand, in the opposing incongruous pairings (i.e., auditory-/ba/ paired with visual-/ga/ and auditory-/pa/ paired with visual-/ka/), fused percepts were much more common and visual responses were extremely rare. This disparity is probably due to the fact that the labial articulation for /p/ and /b/ is more visually salient than that of /k/ and /g/. In a later study by McGurk (1981), adult subjects were presented with auditory /ba/ paired with visual /ba, va, ða, da, za, ga/. In the case of the three most frontal articulations, /ba, va, ða/, the ones with clearly visible articulations, there was complete visual capture (cited in Werker, Frost, & McGurk, 1992). As a result, Werker et al. (1992) state:

> in bimodal speech perception, when the visible articulation – the viseme [...] – unambiguously specifies a particular place of articulation, visual capture can be anticipated. On the other hand, where the viseme is associated with a range of possible places of articulation, visual bias (as shown in "blends") is more likely result. (p. 553)

Indeed, as far as we are aware, high rates of visual capture have never been reported in cases where the place of articulation is not visible.[3] McGurk (1981) reported some instances of visual capture occurring for visual /da/ paired with auditory /ba/, although the fused percept of /va/

[3]Visible articulations generally include labial and dental articulations.

was much more likely. Moreover, we know of no existing study which presents evidence of visual capture occurring in vowels, although the McGurk fusion effect has been shown to take place (e.g., Traunmüller & Öhrström, 2007). Based on observations from previous studies, it seems then that visual capture may be anticipated when the phonetic cues in the visual signal are more perceptually salient than the ones in the acoustic signal (i.e., for certain place of articulation cues in consonants) and when the visual cue unambiguously specifies the phoneme under presentation (i.e., is a viseme), making visual capture with vowels arguably unlikely.

Visual dominance over the auditory modality has been shown to occur in non-speech signals too, indicating that there may be an underlying bias to pay special attention to visual cues more generally. One of the most famous examples is depicted in the Colavita visual dominance effect (Colavita, 1974). The basic experimental paradigm involves a random order of (non-speech) auditory, visual and audio-visual stimuli being presented to subjects who are instructed to make one response whenever they see a visual target and another response whenever they hear an auditory target. For example, participants are instructed to press one button in response to an auditory stimulus and another button in response to a visual one. In the original experiment (Colavita, 1974), participants were not informed that both the auditory and visual stimuli may occur together, while in more recent ones, participants were explicitly told that trials containing both modalities may occur, and in these instances, they should press both the auditory and the visual buttons together (e.g., Koppen & Spence, 2007). Regardless of how informed participants might have been, many studies have shown that while subjects respond to unimodal auditory and visual trials with no problem, they fail to respond to auditory targets when they are presented with auditory and visual targets at the same time (Spence, 2009). Subjects generally respond to bimodal audio-visual tokens with the visual response only. In the original study by Colavita (1974), subjects reported that they had not noticed that the experiment contained bimodal audio-visual tokens as well as unimodal ones. Hecht and Reiner (2009) considered multimodal presentations of various senses including vision, audition and touch. Interestingly, they found the same visual dominance effect in bi-sensory visual-tactile stimuli, but no bias towards either modality in bi-sensory audio-tactile stimuli, suggesting that

dominance may be specifically visual in nature (Spence, 2009).

## 1.4 Visual cues and theories on the objects of speech perception

It is now widely accepted that the perception of speech is influenced by what we see as well as by what we hear. As a result, audio-visual speech perception has played a role in the ongoing debate over the objects of speech perception (Rosenblum, 2008a). Notably, researchers are divided on whether the mechanism for audio-visual integration is innate or whether it develops with linguistic experience. Proponents of gestural accounts of speech perception such as Motor Theory (Liberman & Mattingly, 1985) and Direct Realism (Fowler, 1986) have interpreted audio-visual integration as direct evidence that speech is represented as articulatory gestures (and not sounds). In their view, as speech is underlyingly represented as articulatory gestures, the fact that speech perception is enhanced with the visual cues of these gestures is not surprising (Desjardins et al., 1997; Rosenblum, 2008a). On the other hand, supporters of less controversial auditory-based theories of speech perception (e.g., Diehl & Kluender, 1989; Massaro, 1987; Ohala, 1996; Stevens, 1989) suggest that visual speech input integrates with the acoustic input over the course of development due to increased linguistic experience (Rosenblum, 2008a).

Given children's lack of experience in comparison to adults, one way in which researchers have responded to the question of whether the underlying representation of visual speech requires linguistic experience to develop is to consider the perception of speech in young children and infants (Desjardins et al., 1997). However, as we will show, perceptual evidence from children is mixed and is therefore open to interpretation. Studies have found that pre-linguistic infants less than 7-months-old are sensitive to the correspondence between the auditory and visual speech signals (P. Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Others have suggested that pre-linguistic infants show evidence of the McGurk Effect (Burnham & Dodd, 2004) and may use visual information about speech articulation to learn phoneme boundaries (Teinonen, Aslin, Alku, & Csibra, 2008). These results would therefore support an

integrated, multimodal representation of articulatory and acoustic phonetic information at a very young age (Patterson & Werker, 1999).

However, as we briefly indicated in Section 1.2 (p. 11), a variety of researchers have observed that children are less sensitive to visual speech cues than adults, which would suggest that visual cues may not initially be well specified in children's representations of speech. In the original demonstration of the McGurk Effect, as well as adults, McGurk and Macdonald (1976) also considered the impact of visual speech cues on the perception of children aged 3-5 and 7-8 years. The number of non-auditory percepts (i.e., visual capture, fused and combination responses) was smaller in children than in adults in all stimulus contexts. These results have since been replicated in other studies. For example, Massaro (1984) found that children aged 4-9 years present about half of the visual influence shown by adults in incongruous audio-visual combinations of /ba/ and /da/ and Desjardins et al. (1997) report nearly 60% less visual capture in incongruous audio-visual combinations of /ba, va, da, ða/ in children aged 3-5 years than in adults. The fact that children benefit less from visual cues than adults has also been observed in congruous audio-visual speech. Ross et al. (2011) tested the audio-visual speech recognition abilities in typically developing children aged between 5 and 14 years and compared them to those in adults. They found that children benefited less from observing visual articulations in speech in noise and that this difference tended to be more pronounced as the amount of noise increased. Even children between the ages of 12 and 17 years performed less well than adults. As a result, Ross et al. (2011) concluded that visual enhancement of speech continues to increase until adolescence, and maybe even into adulthood. Finally, Lalonde and Frush Holt (2015) examined developmental differences in the ability to use visually salient speech cues and visual phonological knowledge in 3- and 4-year old typically developing children. They found that visual saliency contributed to audio-visual speech discrimination benefit in all age groups. In a speech recognition task where participants listened to a word presented in noise and were asked to repeat it out loud, 4-year-olds' and adults' substitution errors were more likely to involve visually confusable phonemes in the audio-visual condition than the auditory-only one, suggesting that they used visual phonological representations and knowledge to take

advantage of visually salient speech cues. In contrast, 3-year-olds showed no evidence of this visual phonological knowledge in their substitution errors. As a result, Lalonde and Frush Holt (2015) concluded there may be developmental differences in the mechanisms of audio-visual benefit.

### *1.4.1 The perception-production link*

Given the results from the aforementioned studies, it seems then that even very young infants are sensitive to visual information from speech, but audio-visual speech perception and visual phonological representations take time and linguistic experience to fully form. This is perhaps not that surprising as the same could be said for the development of auditory phonological representations of speech. But what is it about the linguistic experience that makes audio-visual integration possible? Do underlying representations of visual speech emerge from the experience of seeing speech or does experience of producing speech also play a role? This question has been addressed once again by looking at the perception and production of speech in children. Desjardins et al. (1997) tested the hypothesis that young children have not yet had the opportunity to specify fully their representations of visible speech because they have had less experience of correctly producing speech than have adults. They divided a group of 16 4-year-olds into two groups according to whether they made substitution errors or not for the consonants /θ, ð, b, d, v/ in their production. The results indicated that children who substitute are poorer lip-readers and are less influenced by the visual component in incongruous audio-visual syllables (i.e., they report less visual capture) than those who do not substitute. They concluded that the underlying representation of visible speech is mediated by a child's ability to correctly produce consonants. As the authors remarked, their study does not address whether experience of producing speech is actually required for the establishment of an underlying representation that includes visual information. However, Desjardins et al. noted that as very young infants' percepts are influenced by visual speech cues despite not being able to produce consonants themselves, experience of producing consonants cannot be absolutely essential.

While Desjardins et al. (1997) considered the impact of production on perception, other

researchers have considered the impact of perception on production. It has been suggested that access to visual speech cues may aid children to acquire an adult-like articulation of certain speech sounds. It is generally agreed that children produce consonants with observable labial articulations such as /p, b, m/ before non-labial consonants (Steinberg & Sciarini, 2013). Lin and Demuth (2015) presented articulatory data for the acquisition of /l/ in 25 typically developing Australian English-speaking children aged 3;0 to 7;11. Onset /w/ was also included as a control. Lin and Demuth found that children's /w/ productions were dominated by lip rounding, which they argued is due to the visual accessibility of the labial articulation in /w/ productions. In coda /l/, the most common articulation in children was vocalised, i.e., it was produced with a posterior lingual constriction accompanied by a labial constriction. An intermediate articulation between vocalised and adult-like coda /l/ was also observed in which children drop the labial constriction and add or enhance the adult-like lingual constriction. Lin and Demuth speculated that lip rounding may be dropped during acquisition in accordance with visual feedback that a labial constriction is not typical for coda /l/. Visual cues of adult articulations may thus be utilised by children as visible feedback during the acquisition process.

Similarly, in congenitally blind speakers, it has been suggested that a lack of visible speech cues has an impact on both the perception and production of speech. Ménard, Dupont, Baum, and Aubin (2009) investigated the production and perception of Canadian French vowels in blind and sighted speakers and found that while visually-impaired speakers showed greater auditory acuity than sighted speakers, their vowel space is significantly smaller, perhaps due to a reduced magnitude of rounding contrasts. The authors interpreted these results as an indication that the availability of visible speech cues influences speech perception and production. In another study, Ménard, Trudeau-Fisette, Côté, and Turgeon (2016) observed that in clear speech, lip movements were larger in sighted speakers but not in visually impaired speakers, which again indicates that having access to visual cues influences the perception and the production of speech.

## 1.5 Visual cues and spoken language evolution and change

### *1.5.1 Evolution*

Throughout this chapter we have shown that speech perception is influenced by what we see as well as by what we hear. Given the significant impact of visual cues, Rosenblum (2008a) attested that 'multimodal speech is the primary mode of speech perception' (p. 51) and as a result, argued that language must have evolved to be both heard and seen. He pointed out that there should therefore be evidence for the influence of multimodal speech on the evolution of spoken language. While we would rather not enter into the debate on how language evolved, some theorists have argued that the first true language was gestural and not vocal in nature, which may account for the persisting contribution of visual cues to speech perception today. For example, links have been made between the communicative systems in our pre-linguistic ancestors and those in modern-day apes. Corballis (2014) claimed that the closest equivalent to language in nonhuman primates are manual systems because, unlike their vocal calls, their manual gestures are 'intentional and subject to learning' (p. 57). Furthermore, according to Corballis, the fact that much greater success has been achieved in teaching the great apes to speak through gesture rather than vocalisation further indicates that language evolved from manual gestures. In Corballis (2003), he argued that in the evolution of language, vocal elements gradually joined the initial manual gestures, resulting in an association between the two, which provoked the lateralisation of language to the left hemisphere of the brain. The Broca's Area, located in left hemisphere, is predominantly associated with language processing and speech production in humans. The equivalent area in monkeys, however, is more involved in manual action than in humans, but contains the so-called 'mirror' neurons just as it does in humans. These mirror neurons are activated in the brain when a monkey both produces an action, such as grasping a peanut, and when it perceives another individual producing the same action. These mirror neurons thus allow monkeys to understand gestural action. According to Corballis, the presence of these mirror neurons in our pre-linguistic ancestors may have set the stage for the evolution of language. Vocalisation must have been incorporated into the mirror system,

which was initially specialised for manual grasping but became increasingly differentiated and lateralised for language in the course of human evolution. This also explains why language may be both vocal, as in spoken languages, or manual, as in signed languages. Corballis also proposed that it is this lateralisation to the left hemisphere which may be responsible for the uniquely human trait of right-handedness today, which although speculative, could be seen as phylogenetic evidence for the evolutionary basis of multimodal speech.

As Rosenblum (2008a) remarked, other evidence for the multimodal basis of the evolution of speech may lie in the phonological inventories of the world's languages. Indeed, as cited in Section 1.1 (p. 8), Summerfield (1983) suggested there is a 'fortunate complementary relationship between what is lost in noise or impairment, and what can be provided by vision' (p. 183). We may wonder then if this 'fortunate' relationship has occurred because language evolved to ensure that sounds that are hard to hear are easy to see, and vice versa, and indeed Rosenblum (2008a) asks the same question. He hypothesised:

> If visual speech does constrain phonological inventories, the world's languages should include relatively few phonetic segments that are both difficult to hear and *see*. (p. 67, emphasis original)

As we have already observed, phonetic contrasts that are difficult to hear typically involve place contrasts in consonants, such as /p/ versus /t/ and /m/ versus /n/. Incidentally, Dohen (2009) noted that the visible salience of the latter /m/-/n/ contrast may explain why it exists in almost all of the world's languages, which would follow the argument that the complementarity of audio-visual speech cues is no accident. However, more research is required which specifically considers this question and accounts for the phonological inventories of many languages, including those which are underrepresented in the literature.

We would like to stress that we do not mean to say that the multimodal nature of speech perception can exclusively explain why phonological inventories have evolved in the way that they have. For example, Stevens's Quantal Theory (1989) suggests that the most frequent sounds in the world's languages may be accounted for by considering the nature of articulatory-acoustic relations within the human vocal tract. Quantal Theory, which we will revisit later

in Chapter 3 (Section 3.5, p. 84), proposes the existence of 'quantal' areas in the vocal tract, which are associated with acoustic stability. Large changes to the location of a constriction positioned in a quantal region would result in comparatively fewer changes to the resulting acoustics. Conversely, in non-quantal regions, very small articulatory deviations would result in large changes from one acoustic output to another. Given the acoustic stability resulting from articulations in quantal regions, Quantal Theory would argue that the most frequent sounds are produced in these regions where speakers have the most articulatory freedom.

### *1.5.2 Sound change*

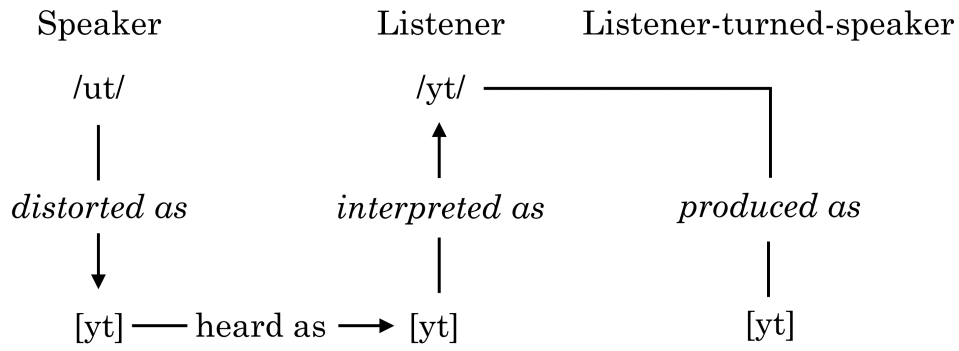
We propose that Rosenblum's (2008a) evolutionary account of multimodal speech perception could also be extended to diachronic sound change. If the perception of speech is inherently multimodal, visual cues may also be implicated in sound change. However, models of sound change tend to neglect the role of visual cues and focus instead on the impact of the auditory ones. While the most well-known models of sound change are divided on who initiates sound change, the listener or the speaker, they generally converge on the notion that speech is perceived from the auditory signal alone. In this section, we will briefly describe two of the most well-known models of sound change, categorised as perception-oriented (i.e., listener-based) and production-oriented (i.e., speaker-based) accounts, and will present empirical evidence which suggests that visual cues may well play a part in sound change and in the shaping of the sound systems of the world's languages.

The most famous perception-oriented account of sound change is provided by Ohala, who asserts that the main source of variation in speech, and hence the driving force behind sound change, is the misperception of the acoustic signal by the listener (e.g., Ohala, 1981). In his view, much of the variation which underpins the acoustic speech signal is phonetically predictable. When the phonetically experienced listener is able to factor out this variation, sound change does not occur. In contrast, sound change can be triggered when the listener takes the acoustic signal at face value and fails to apply their phonetic knowledge of how speech sounds interact in perception (Chitoran, 2012). When the listener turns speaker, he may thus produce a new

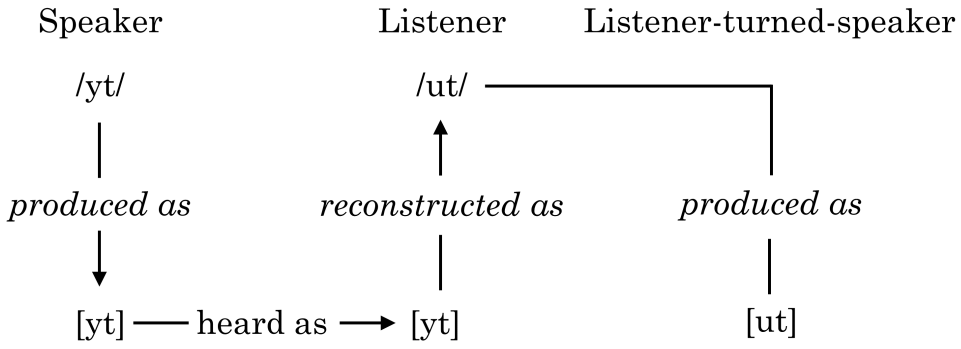
form, which is different to the one intended by the original speaker, which Ohala termed *hypocorrection*. Another scenario which may result in sound change, labelled *hypercorrection*, occurs when the listener performs an erroneous correction of the acoustic speech signal, again resulting in a new form in his/her production. Ohala (1981) provided the example of the vowel /u/, which may be subject to assimilation in the context of a surrounding anterior consonant such as /t/, e.g., /ut/ may have the surface form [yt]. In the case of hypocorrection, as schematised in Figure 1.2a, the listener fails to reconstruct [yt] as the intended /ut/, which is interpreted as /yt/ and then, when the listener turns speaker, is produced as [yt], triggering a sound change. In the case of hypercorrection, schematised in Figure 1.2b, the speaker intends to produce /yt/ and does so appropriately, resulting in the surface form [yt]. The listener incorrectly reconstructs the intended /yt/ as /ut/ given his phonetic knowledge of assimilation in this particular context, which results in a production of [ut] when it is the listener's turn to speak. As Chitoran (2012) pointed out, both of these scenarios imply a mismatch between production and perception in the listener.

Production-oriented accounts of sound change, notably the one proposed by Lindblom (1990), converge with perception-oriented ones in that they too consider phonetic variation in speech to be the impetus for sound change. However, the source of this variation is considered to originate from the speaker as opposed to the listener. In his 'Hyper'- and 'Hypo'-articulation (H&H) Theory[4], Lindblom proposes that speech varies on a continuum from hyperarticulated listener-oriented *clear speech* to hypoarticulated speaker-oriented *casual speech*. The speaker's aim is to produce utterances that are intelligible to the listener, but to do so expending as little energy as possible. As J. F. Hay, Sato, Coren, Moran, and Diehl (2006) noted, speakers try to achieve sufficient, as opposed to maximal, distinctiveness in their articulation of speech sounds, and thus make active adjustments to their production of speech according to the predicted perceptual needs of the listener and to their own articulatory needs. In hyperarticulated speech, the listener's perceptual needs take precedence over the speaker's articulatory needs, which requires more effort from the part of the speaker. In hypoarticulated speech, the speaker uses

[4]H&H Theory will be revisited later in the thesis, notably in Experiment 1 when we discuss hyperarticulation in more detail.




**(a)** *Hypocorrection*




**(b)** *Hypercorrection*

**Figure 1.2:** *Listener-oriented sound change scenarios according to Ohala (1981) including* **(a)** *hypocorrection and* **(b)** *hypercorrection.*

minimal articulatory effort to conserve energy but the listener's perception may suffer as a consequence. Hyperarticulation is therefore at odds with hypoarticulation: hyperarticulation increases perceptibility in the listener, but requires additional effort from the speaker. Sound change is therefore goal-driven (i.e., teleological) and predicted to arise when a speaker feels the need to adjust their articulation to one which is either easier to perceive or easier to produce.

Although the models proposed by Ohala and Lindblom do not converge on who is the initiator of sound change, the listener or the speaker, speech perception is viewed by both models as the transformation of the *auditory* input signal into neural representations of speech sounds in the listener. Other modalities involved in the perception of speech such as visual cues are notably absent from both models. Up to this point we have considered how sound change is modelled in phonological theory. Like many good theories, these approaches have been built on extensive experimental work, as Chitoran (2012) noted. We now need to consider whether incorporating visual cues to sound change models is actually necessary, based on empirical evidence from the literature. We will present two cases from English which suggest that visual speech cues may indeed be implicated in sound change. This evidence demonstrates the need to consider visual as well as auditory speech perception in sound change models.

The phonetic realisation of the /f/-/θ/ contrast in English is well-known for being acoustically ambiguous. In acoustic terms, [f]-[θ] lack spectral peaks and have very low intensity, which makes them difficult to differentiate (Tabain, 1998). In native speakers of English, [θ] is regularly fronted to [f], particularly in British accents. Listener-driven models of sound change would explain the change from /θ/ to /f/ as the misperception of [θ] in the listener, given its acoustic similarity to [f]. However, McGuire and Babel (2012) noted that listener-driven models cannot account for the fact that while the sound change from /θ/ to /f/ is widely attested cross-linguistically, there are no known cases of /f/ being substituted for /θ/ in the literature on language typology.[5] McGuire and Babel (2012) therefore described an 'asymmetry' in the /f/-/θ/ substitution pattern. They proposed that a bias towards /f/ originates in the

[5]Interdental fricatives are also typologically rare more generally. Only 7% of the 451 languages included in the UCLA Phonological Segment Inventory Database (UPSID) show interdental fricatives (Maddieson, 1984; Maddieson & Precoda, 1989).

greater visual saliency and stability of /f/. As McGuire and Babel noted, it has been remarked in previous studies that the visual cue of the lips may be more informative than the acoustic cues in disambiguating /θ/ and /f/ (e.g., Jongman, Wang, & Kim, 2003; Miller & Nicely, 1955). McGuire and Babel (2012) considered how visual information may be implicated in the sound change involving /θ/ and /f/ by examining the role of visual cues in the perception of the contrast across multiple speakers in American English. Their results suggested that /θ/ is more variable than /f/ in both articulation and acoustics. For example, the visibility of the tongue gesture for /θ/ varied across the speakers who served as perception stimuli because it was produced both inter-dentally and dentally. Furthermore, the acoustics of /θ/ in the same speakers substantially differed across different vowel environments. McGuire and Babel therefore proposed that it is this variability which has contributed to the unstable nature of /θ/ across time, which they argued offers an explanation for the asymmetry in the patterning of /f/ and /θ/. In their view, listeners are faced with unpredictable inter-speaker variability in the production of /θ/ and failure to perceive either an auditory or a visual /θ/cue will lead to the sound being categorised as /f/ based of their acoustic and visual phonetic similarities. As a result, McGuire and Babel concluded that their results demonstrate the need to consider multimodal phonetic information when theorising about sound change, as well as in discussions on acquisition and on typological distributions of sounds in the world's languages.

Another acoustically ambiguous contrast involves the /ɔ/-/ɑ/ contrast in certain varieties of American English due to the Northern Cities Vowel Shift, in which both vowels undergo fronting, resulting in a merger. Havenhill and Do (2018), which presented work from Havenhill's thesis (2018), considered both the production and the perception of the /ɔ/-/ɑ/ contrast in American English. Articulatory data indicated that some speakers distinguish /ɔ/ from /ɑ/ with a combination of tongue position and lip rounding, while others used either tongue position or lip rounding alone, which has acoustic consequences: /ɑ/ and /ɔ/ are more similar in the cases in which only one articulatory dimension varies, as opposed to two. While all speakers maintained some degree of acoustic contrast between the vowels, Havenhill and Do considered the impact of visual cues to the perception of the /ɑ/-/ɔ/ contrast. They found

that despite having a similar acoustic output, the articulatory configurations in which /ɔ/ is produced with unrounded lips are perceptually weaker than those produced with visible rounding. Unrounded /ɔ/ was more likely to be (mis)perceived as /ɑ/ than rounded /ɔ/ when listeners had access to visual speech cues. Havenhill and Do argued that their results showed that visual cues may play a role in shaping phonological systems through misperception-based sound change. They proposed that visual speech cues can inhibit misperception of the speech signal in cases where two sounds are acoustically similar, which suggests that phonological systems may be 'optimised' for both auditory and visual perceptibility. Like McGuire and Babel (2012), Havenhill and Do (2018) also concluded that theories on language variation and sound change must consider how speech is conveyed across multiple perceptual modalities.

## 1.6 CHAPTER CONCLUSION

In this chapter, we have shown that speech perception is inherently multimodal in nature. Having access to visible speech cues can not only enhance perception, particularly when the auditory conditions are degraded, but in some cases, may actually influence or override the auditory perception of speech. The most dramatic demonstration of the influence of visual cues arguably occurs in incongruous audio-visual speech perception in the laboratory, when listeners are 'visually captured', i.e., they report hearing the sound they saw, as opposed to the sound they actually heard. However, there is a wealth of evidence to suggest that speakers use visual cues well outside of the laboratory in their everyday lives. For instance, visual cues may provide children acquiring language with articulatory feedback, which may help them to reach adult-like articulations. Furthermore, speakers produce more visually intelligible speech cues in clear, or hyperarticulated speech, which suggests they are aware of the benefits of producing visually salient phonetic cues in increasing speech intelligibility. This behaviour likely develops with experience of seeing speech because blind speakers tend not to enhance speech visually. Whether linguistic experience is required for audio-visual speech integration is a subject of much debate. Although it is generally agreed that children benefit less from visual speech cues than adults and that visual enhancement of speech perception appears to develop

throughout childhood and adolescence, pre-linguistic infants do show sensitivity to visual speech cues. As we have observed, it has been suggested that language may have evolved to be both heard and seen and that the first true language may have been gestural in nature. There may even be phylogenetic evidence for the evolution of speech as a multimodal audio-visual entity. Phonological inventories of the world's languages may have evolved to some extent to ensure that sounds that are hard to hear are easy to see and this evolutionary process may well still be ongoing in sound change.

# The complex articulation of English approximant /r/ 2

## 2.1 Why English /r/?

The phonetic implementation of the English approximant consonant /r/ has been described as one of the most complex articulations in the English language (Adler-Bock, Bernhardt, Gick, & Bacsfalvi, 2007). What makes the articulation of this sound so unique is the variability it entails, particularly with regards to the realisation of the palatal constriction, which has been extensively studied in the Englishes of North America and Scotland. A variety of imaging techniques have been employed to observe the articulation of /r/ in these varieties including cineradiography (x-ray films) (e.g., Delattre & Freeman, 1968; Westbury, Hashi, & Lindstrom, 1998; Zawadzki & Kuehn, 1980), Electromagnetic Articulography (EMA) (e.g., Boyce & Espy-Wilson, 1997; Guenther et al., 1999), lingual probe contact (Hagiwara, 1995), magnetic resonance imaging (MRI) (e.g., Alwan, Narayanan, & Haker, 1997; Boyce, Tiede, Espy-Wilson, & Groves-Wright, 2015; Proctor et al., 2019; Zhou et al., 2008), and Ultrasound Tongue Imaging (UTI) (e.g., Heyne, Wang, Derrick, Dorreen, & Watson, 2018; Lawson et al., 2013; Mielke, Baker, & Archangeli, 2016). The numerous articulatory studies on English /r/ have shown that the post-alveolar approximant may be produced with a number of different tongue body shapes,

which are categorised on a continuum between two extreme configurations: bunched and retroflex. The wide range of articulatory variation associated with /r/ is generally considered covert, meaning that the differences across speakers and syllable contexts are not perceptually salient. This covert articulation makes /r/ a particularly interesting case from which to test theories on speech production and has thus provided the grist for important work on the link between articulation and acoustics, and on individual variation.

Despite the array of existing articulatory studies, the production of /r/ still remains enigmatic, especially with regards to the secondary articulations which accompany the lingual component. /r/ typically involves three simultaneous constrictions in the vocal tract: in the pharynx, in the mid-palatal region and at the lips (Espy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 2000; Fant, 1960; Westbury et al., 1998). It is the latter labial constriction which is of particular interest in this thesis. It is generally agreed that /r/ may be labialised, particularly in pre-vocalic and pre-stress syllable positions in both American English (Delattre & Freeman, 1968; Mielke et al., 2016; Proctor et al., 2019; Uldall, 1958; Zawadzki & Kuehn, 1980) and the variety of English spoken in England, henceforth Anglo-English (Abercrombie, 1967; Jones, 1972; Scobbie, 2006). However, the exact contribution of the lips to English /r/ has yet to be explored in any variety of English, which, as Docherty and Foulkes (2001) noted, may have resulted in a 'skewed view of the physical basis' of /r/ (pp. 182-183).

The lips may have a particularly important contribution to the production of /r/ in Anglo-English, as labiodental variants are gaining currency (Docherty & Foulkes, 2001; Marsden, 2006). It is generally implied that labiodental variants have emerged in England by speakers retaining the labial gesture of /r/ at the expense of the lingual one (Docherty & Foulkes, 2001; Foulkes & Docherty, 2000; Jones, 1972), perhaps due to the heavy visual prominence of the lips (Docherty & Foulkes, 2001). However, the articulation of /r/ in Anglo-English has received very little empirical attention, which may be due to its non-rhotic status. /r/ is an important and well-known sociolinguistic marker, dividing English varieties into rhotic and non-rhotic. While rhotic accents pronounce all orthographic 'r's, non-rhotic ones only allow /r/ to be produced when directly followed by a vowel. Rhotic accents include the typical accents of most

of the United States, Canada, Scotland, Ireland and Barbados. Meanwhile, the typical accents of Australia, New Zealand, South Africa, Trinidad, certain eastern and southern parts of the United States, and most of England and Wales are considered non-rhotic (Wells, 1982). Indeed, although empirical accounts of the production of /r/ in Anglo-English are few and far between, the contexts in which /r/ may be produced in this variety have solicited a lot of attention, particularly with regards to the phonological process of hiatus-filling /r/-sandhi, which will be briefly discussed in Section 2.3 (p. 39). The somewhat misleading terminology, *non-rhotic*, may be to blame for the previous lack of interest in the articulation of /r/ in Anglo-English but non-rhoticity does not imply that /r/ is not pronounced at all. As Carr and Durand (2004) observed, it is simply the presence of /r/ in syllable codas which defines an accent as rhotic, and not the quality of the 'r'. However, given the lack of articulatory studies of /r/ in non-rhotic varieties, we cannot be sure that the phonetic quality of prevocalic /r/ is the same across Englishes. Indeed, despite the lack of empirical data, Anglo-English /r/ tends to be more associated with tip up tongue postures than tip down ones, contrary to /r/ in American English, in which tip down bunched shapes are more frequent (e.g., Delattre & Freeman, 1968). There is therefore a perception among some phoneticians that the articulation of Anglo-English /r/ is not as variable as that of rhotic Englishes, which may also account for the apparent lack of interest in the production of /r/ in England. Indeed, at the most recent 6$^{th}$ edition of the *R-atics Colloquium* in Paris, the international conference dedicated to the study of 'r'-sounds, one researcher made the informal observation that if we were to make a list of all the variability in the world's 'r'-sounds, the Standard Southern British English (SSBE) post-alveolar approximant would surely be at the very bottom of the list. Yet, no large-scale articulatory study of the /r/ produced in England currently exists.

## 2.2 Defining Anglo-English

This thesis will therefore present data from the non-rhotic variety of English spoken in England, Anglo-English, which has been understudied with regards to the phonetic implementation of its post-alveolar /r/. We note that the term *Anglo-English* was chosen rather than the traditional

SSBE label because we do not focus exclusively on the Standard Southern variety. The data we will present come from speakers from all over England who do not necessarily use SSBE. We use the term Anglo-English rather than British English to avoid confusion with other varieties of English spoken in Scotland, Wales and Northern Ireland. Although remnants of rhoticity still exist in some pockets of England, such as the North West and South West (Wells, 1982), derhoticisation is nearly complete. Piercy (2012) analysed the status of rhoticity in 24 speakers from the south west of England aged between 14-83 years at the time of data collection. She found a significant correlation between rhoticity and age: as age decreases the use of non-prevocalic /r/ decreases. Rhoticity was absent in all participants under the age of 30, indicating that the change from rhotic to non-rhotic is complete in this region of England. Perhaps the most revealing data of recent times come from the *English Dialects App*, in which users indicate which variants of 26 words they use and the application 'guesses' their local dialect (Leemann et al., 2018). By asking participants if they pronounced the 'r' sound in the word *arm*, the rhotic status of each user can be judged. Leemann et al. (2018) present the results on rhoticity from roughly 29 000 respondents from the UK and Ireland and compare their geographical distribution with those from the *Survey of English Dialects* (Orton & Dieth, 1962), which was collected between 1950 and 1961 in 313 localities across England. Figure 2.1 from Leemann et al. (2018) reveals a striking trend for non-rhoticity (in green) in the 2016 data, which was much less widespread in the 1950s. The geographical distribution from 2016 follows results from previous studies in that remnants of rhoticity now only remain in the south and north west of England, but no area presents more than 45% rhoticity among the people surveyed.

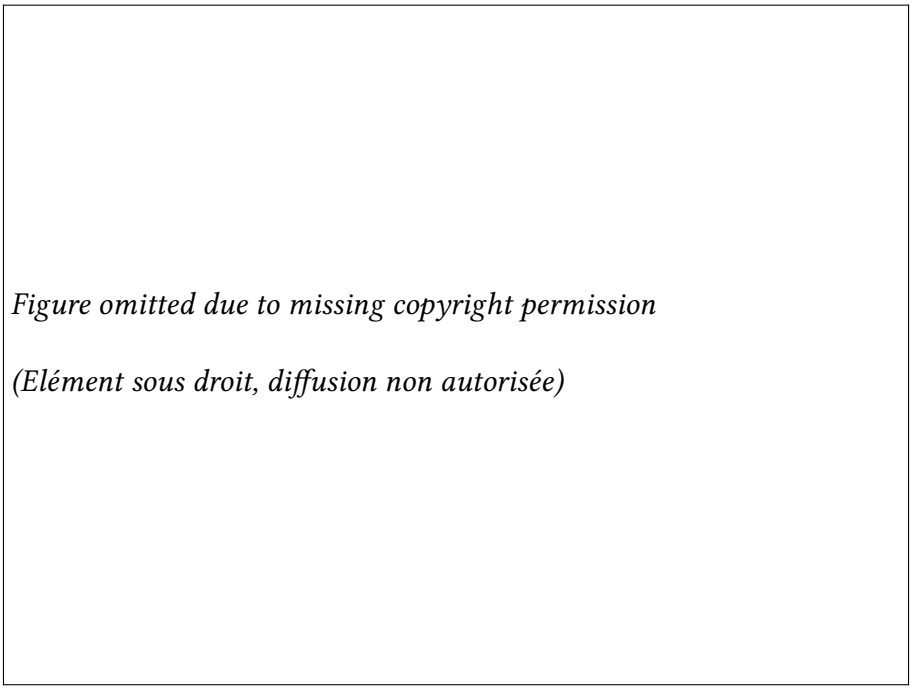



**Figure 2.1:** *The geographical distribution of rhoticity based on data from the* Survey of English Dialects *from the 1950s (left) (Orton & Dieth, 1962) and the* English Dialects App *from 2016 (right) (from Leemann et al., 2018, p. 12).*

## 2.3 PHONOLOGICAL ASPECTS OF RHOTICITY

Within the studies on English rhoticity, a large body of work has studied the phonological process of hiatus-filling /r/-sandhi. In fact, when it comes to Anglo-English /r/, most of the work focuses, not on how /r/ is produced, but when. It is generally assumed that rhoticity and /r/-sandhi are in 'complementary distribution' (Barras, 2008) and Giegerich (1999) even went as far as to suggest that /r/-sandhi is 'systematically confined' to non-rhotic English (p. 168). In non-rhotic accents, /r/ is only pronounced when directly followed by a vowel. As a result, in words which end in an etymological and orthographic /r/, i.e., the word *car*, there is generally an alternation in pronunciation: [ɹ] before a vowel, ∅ elsewhere (Foulkes, 1997). When these words which end in /r/ occur before words beginning with vowels, /r/ is pronounced. For example, *car driver* would be pronounced [kɑː ˈdɹaɪvə] but *car and driver* would be pronounced [kɑːɹ ən ˈdɹaɪvə]. This phenomenon is known as *linking* /r/. In many linking /r/ users, the

process has been extended to words which do not historically or orthographically contain /ɹ/, termed *intrusive* /ɹ/, as in *saw it* which may be pronounced [sɔːɹ ɪt]. According to most phonological accounts of /ɹ/-sandhi, rhotic speakers would always pronounce the /ɹ/ in *sore* but never in *saw*. However, this assumption has been challenged as intrusive /ɹ/ has been observed in archive recordings of rhotic New Zealand English from the 1940s (J. Hay & Sudbury, 2005), and in more modern recordings of speakers from the north west (Barras, 2010) and the south west (Werner, 2019) of England, where rhoticity is still present in some speakers. Although these studies suggest that rhoticity and /ɹ/-sandhi are not necessarily mutually exclusive, there is a general tendency that the more rhotic a speaker is, the less likely they are to produce intrusive /ɹ/. As Barras (2010) noted in his PhD thesis, 'the overall picture matches general historical accounts of the emergence of /ɹ/-sandhi, in which it is claimed to develop after a loss of rhoticity' (p. 265).

Indeed, the rhotic status of a variety of English is not necessarily as black and white as what it might seem on the face of it. In some varieties, the status of rhoticity is shifting. For example, the non-rhotic accents of the United States are becoming increasingly rhotic (Labov, Ash, & Boberg, 2008), while a process of derhoticisation is underway in working class speakers of Scottish English (Stuart-Smith, 2007). Some varieties of English may even be considered 'hyper-rhotic' or 'hyper-dialectal' in that /ɹ/ occurs in non-etymological, non-sandhi environments, i.e., utterance finally or in coda consonant clusters (Barras, 2010). For example, utterance final *idea* may be pronounced [aiˈdɪəɹ]. In England, hyper-rhoticity is considered a feature of traditionally rhotic dialects due to contact with non-rhotic varieties (Trudgill, 1986; Wells, 1982), particularly in the south west (Barras, 2010). Conversely, in North America, hyper-rhoticity is associated with varieties which were once non-rhotic but have become rhotic due to pressure from General American (Krämer, 2012). It is interesting that hyper-rhoticity is due to contact with both rhotic and non-rhotic variants in America and England, respectively. Accommodation Theory (H. Giles & Smith, 1979) may give a sociolinguistic explanation for this somewhat paradoxical observation. The theory states that speakers make modifications to their accent in order to converge to or diverge from those of their interlocutors, depending on whether

they wish to identify with or distance themselves from those interlocutors (Beal, 2009). While hyper-rhoticity may be considered convergence to the standard non-rhotic variant in North America, in the south west of England, rhotic speakers may be diverging from non-rhotic SSBE with hyper-rhoticity. Regardless of whether Accommodation Theory can accurately account for the development of hyper-rhoticity across Englishes, it is safe to say that dialect contact plays a role in shifts in rhoticity and that the situation is not necessarily as straightforward as what the labels *rhotic* and *non-rhotic* would imply.[1]

## 2.4 TONGUE SHAPE DIVERSITY

Without a doubt, the most widely studied aspect of the articulation of English /r/ is its lingual component. It was originally thought that there were two distinct tongue shapes for American English /r/. For example, from studying palatograms of her own production of /r/, Uldall (1958) described two configurations: 'molar' and 'tongue tip'. She defined molar /r/ as a bunching of the tongue towards the upper back molars and the drawing back of the tongue tip into the tongue body. Tongue tip /r/ was characterised by Uldall as the raising of the tongue tip behind the alveolar ridge with the front of the tongue held concave to the palate. Over the years, many more tongue shapes have been described, but it is generally agreed that Uldall's 'molar' and 'tongue tip' /r/, now more commonly referred to as bunched and retroflex, exhibit the greatest degree of contrast (Zhou et al., 2008).

The well-cited Delattre and Freeman (1968) cineradiographic study was the first to indicate that rather than two possible variants, tongue postures for /r/ should be considered to be on a continuum of possible shapes with bunched /r/, whose primary constriction occurs at the tongue dorsum, and retroflex /r/, whose primary constriction occurs at the tongue tip, at its endpoints. The continuous nature of tongue shapes has since been corroborated by more recent

[1] Indeed, rhoticity may even fluctuate across the lifespan depending on an individual speaker's input. Anecdotally, changes to rhoticity occurred in a family member who moved to Scotland as a young child from the north of England. She not only acquired rhoticity, but produced hyper-rhotic /r/. For example, her sister's name, *Lydia*, was often pronounced ['lɪdiə̯ɹ]. Hyper-rhoticity gradually eroded when she moved back to England in her teenage years.

studies on American English and Scottish English such as Alwan et al. (1997); Lawson, Scobbie, and Stuart-Smith (2011); Mielke et al. (2016); Tiede, Boyce, Holland, and Choe (2004); Westbury et al. (1998), among others. Delattre and Freeman (1968) recorded 43 American English and 3 Anglo-English speakers and observed eight tongue postures for /r/, as presented in Figure 2.2.[2] Types 1 & 2 were observed postvocalically in non-rhotic speakers, i.e., in 'r'-less contexts, with Type 1 occurring only in English participants. The remaining tongue shapes (Types 3-8) are ordered incrementally from most bunched (Type 3) to most retroflex (Type 8).

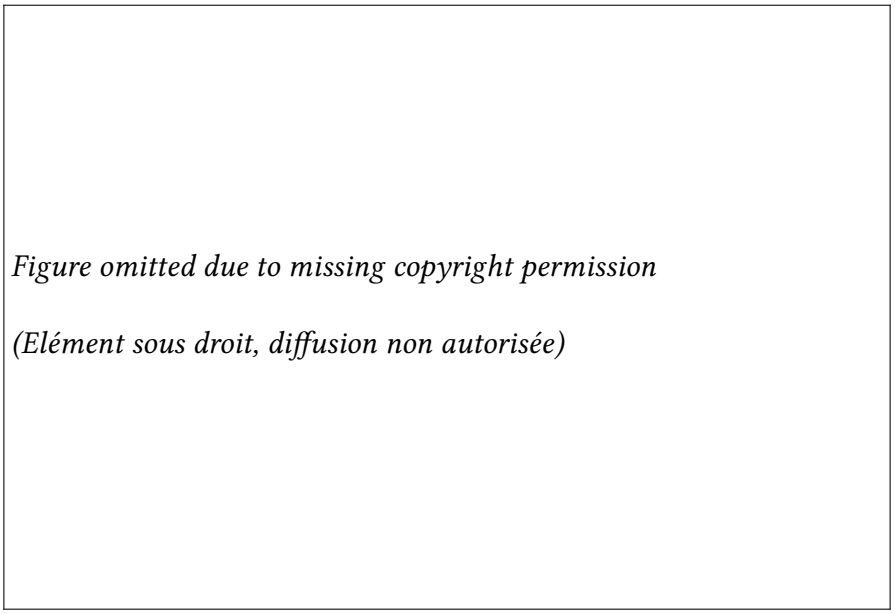



**Figure 2.2:** *Delattre and Freeman (1968)'s taxonomy of tongue shapes for American English and Anglo-English* /r/ *(from Mielke et al., 2016, p. 103).*

An important finding from Delattre and Freeman (1968), which contradicts prior accounts of English /r/, is the fact that the primary tongue shape in American English is bunched and not retroflex. Among the 43 American English speakers presented in Delattre and Freeman (1968), 67 % of tokens were produced with bunched tongue shapes (Types 3-5) across all contexts (i.e., word-initial, postconsonantal, intervocalic pre-stress, intervocalic post-stress, preconsonantal and word-final). This finding has since been confirmed by a variety of articulatory studies

[2] Figure 2.2 was adapted by Mielke et al. (2016) to conform with conventions for the orientation of midsagittal ultrasound images, i.e., with the tongue tip on the right.

on American English /ɹ/ in speakers from so many different regions that the findings seem to be generalisable to all rhotic American Englishes, including Zawadzki and Kuehn (1980) (2/3 exclusive bunchers), Lindau (1985) (6/6 exclusive bunchers), Mielke et al. (2016) (16/27 exclusive bunchers), S. Chen, Tiede, and Whalen (2017) (7/9 exclusive bunchers), and all 24 speakers presented in Bakst (2016) bunched their tongue for /ɹ/, although a small majority retroflexed too. To the best of our knowledge, one study was an exception and found more exclusively tip up users (9/15 speakers) than bunchers (Hagiwara, 1995). This discrepancy may be due limitations in the technique employed, as tongue shapes were inferred by using a cotton swab which was inserted into the mouth to determine whether contact was made on the surface, underside or tip of the tongue for /ɹ/. Despite the abundance of studies indicating that bunching is the most common tongue shape in American English, in early phonetic work prior to the 1960s, the retroflex tongue shape was generally considered the primary articulation of the palatal constriction for /ɹ/. For example, Heffner (1950) indicated that retroflexion of the apex of the tongue was the normal initial /ɹ/ in many American English speakers (as cited in Delattre & Freeman, 1968).

Similarly, classic descriptions of English pronunciation based on SSBE or Received Pronunciation generally converge on the suggestion that English people use a retroflex, or at the very least a tip up, tongue configuration for their prevocalic /ɹ/. Descriptions as early as Sweet (1877) refer to tip up articulations as opposed to tip down ones. Sweet (1877) described the tongue tip pointing upwards and a 'tendency to make the outer front of the tongue concave' (p. 37), which presumably refers to a curled up retroflex articulation. Jones (1972) described the sound of the /ɹ/ as 'the equivalent to a weakly pronounced retroflexed ə' (p. 206). Although Gimson (1980) suggested that the degree of retroflexion may be greater in American English and in the rhotic accents of the south west of England, he did indeed describe it as a retroflex in Received Pronunciation. On the other hand, Ladefoged and Disner (2012) argued that many 'BBC English speakers' use tongue tip raising towards the alveolar ridge, while many American English speakers bunch the body of the tongue up. Indeed, the three Anglo-English speakers presented in Delattre and Freeman (1968) used an 'extreme' tip up shape prevocalically, which

differed from American English shapes.

Interestingly, bunched /r/ is rarely, if ever, mentioned as an alternative strategy in pronunciation manuals for second language learners of English, particularly in those based on SSBE.[3] These manuals strongly focus on retroflexion, encouraging learners to curl the tongue tip back and often provide stylised midsagittal drawings indicating retroflexion (e.g., Ashton & Shepherd, 2012; Hancock, 2003; Marks, 2007; Roach, 1983; Underhill, 1994). Drawing on their experiences as voice and dialect coaches of British English, Ashton and Shepherd (2012) went as far as to suggest that the 'correct position' to produce the /r/ sound in English is with the tongue tip curled back and upwards towards the roof of the mouth (p. 48). Despite the abundance of retroflex descriptions in the literature, a recent small-scale articulatory study of Anglo-English indicated that speakers present similar articulatory variation to speakers with rhotic accents (Lindley & Lawson, 2016) including bunched tongue shapes. But more extensive research is required for a more robust description of the tongue shapes employed for the production of Anglo-English /r/.

Retroflexion has traditionally been described as an articulation involving the curling up of the tongue tip (e.g., Catford, 1977). Yet even in early work, it was reported that the degree of retroflexion may vary in the production of English /r/. For example, Kenyon (1940) observed that the apex of the tongue may merely be raised towards the alveolar region rather than curling back (as cited in Delattre & Freeman, 1968). Delattre and Freeman (1968) also observed retroflex articulations which differed in their degree of curling back of the tongue tip. Type 8 is described as having the tongue tip curled up, which can be seen in the tracings in Figure 2.2 (p. 42). 'Real retroflex' articulations such as Type 8 have been given the label *sublaminal* in order to distinguish their articulation, in which the underside of the tongue blade is curled back over the tongue, from that of *apical* retroflexes, in which the constriction is formed with the tongue apex (e.g. Hagiwara, 1995). An example of an apical retroflex is presented in Type 7 in Figure 2.2, in which the angle of the tongue tip is markedly different from that of Type 8, as discussed in

[3] We found one mention of bunching in a teachers' manual on American English pronunciation (Ehrlich & Avery, 2013). The authors indicated that although there is a 'disagreement' regarding the characterisation of /r/ as either retroflex or bunched, which may be due to 'dialectal differences', they stressed that retroflexion is the most useful characterisation for pedagogical purposes.

Mielke et al. (2016). From Delattre and Freeman (1968)'s taxonomy (as presented in Figure 2.2), Type 8, with its curled up tongue tip, appears to present a sublaminal articulation. However, according to Delattre and Freeman (1968) the main difference between Type 7 and Type 8 is the degree of pharyngeal constriction: unlike Type 7, Type 8 does not have a constriction at the tongue root. Another major difference between Delattre and Freeman's Type 7 and Type 8 lies in the origins of the speakers who produce them. Type 8 is described as 'the strong British /r/ used in prevocalic positions' (Delattre & Freeman, 1968, p. 45) regardless of stress and was not found in any of Delattre and Freeman's 43 American participants. It was therefore suggested that Anglo-English /r/ is produced primarily with retroflexion, unlike American English /r/.

Although Delattre and Freeman (1968) observed that bunched tongue shapes are more common than retroflex ones in American English speakers, retroflex shapes do occur, although they are less extreme than the retroflex tongue posture used by English participants (i.e., Type 8). Types 6 and 7 are considered to be American counterparts of the British Type 8 because they are also found in 'strong syllabic position' (Delattre & Freeman, 1968, p. 46), i.e., initial prevocalic. As previously discussed, Type 7 is described as having an apical articulation. Although Type 6 is labelled *fronted bunched* by Delattre and Freeman, it is still considered a retroflex tongue configuration but with a bladal place of articulation. It only occurs in speakers who use a bunched tongue shape postvocalically and is described as a fronted version of their bunched postures when /r/ occurs prevocalically, i.e., a 'compromise between bunched and retroflex' (Delattre & Freeman, 1968, p. 56). The status of bladal articulations such as this has been disputed with regards to where they are situated on the bunched-retroflex continuum. The problem perhaps lies in the definition of retroflexion. It has been widely reported that the tongue tip may fail to curl up in other languages with segments traditionally considered retroflex (Hamann, 2003). As a result, in her PhD, Hamann (2003) refined the definition of retroflexion and proposed the combination of four articulatory characteristics: apicality, posteriority, sublingual cavity, and retraction. As such, by her definition, any sound articulated with the tongue tip positioned behind the alveolar region, creating a space underneath the tongue, and with a displacement of the tongue back towards the pharynx or velum may be considered retroflex. Given their lack

of apicality, bladal articulations of English /r/ would not necessarily comply with Hamann's definition. Indeed, Hagiwara (1995) and Mielke et al. (2016) both proposed that Delattre and Freeman's Type 6 is more similar to the extreme tip down bunched than the extreme sublaminal retroflex. As Hagiwara noted:

> There is little about the 'blade up' tongue shape which is suggestive of what is normally meant by 'retroflexion'. There is some further, indirect evidence that 'blade up' should not be classed with the other truly retroflex articulations. (Hagiwara, 1995, p. 100)

Hagiwara's 'further, indirect evidence' came from the contextual distribution of tongue shapes in speakers who use more than one shape. In Hagiwara's probe contact American English data, speakers who presented one tongue shape exclusively always used tip up postures, while in other speakers, tip down and blade up shapes were used in combination. According to Hagiwara (1995), this pattern indicates that the blade up and tip down shapes form a class of their own, which is distinct from the alternative tip up class. The same pattern has been observed much more recently by Proctor et al. (2019) who recorded four American English speakers using MRI. They noted that speakers with blade up initial /r/ typically realise syllabic and final /r/ with tip down postures. Delattre and Freeman's articulatory data also followed a similar pattern. They observed a general tendency for speakers who used bunched shapes (i.e., Types 3-5) in postvocalic position to use the bladal Type 6 in prevocalic position. However, Delattre and Freeman (1968) found more retroflexion in prevocalic position than bunching and therefore considered Type 6 to be closer to retroflex than bunched shapes.

Given the disparities in accounts concerning bladal articulations of /r/, a solution may be to go beyond the dichotomous bunched-retroflex classification. Such a strategy was employed by Espy-Wilson et al. (2000), who used a three-way categorisation: tip up retroflex; tip up bunched; tip down bunched. The bladal configuration would therefore be considered a tip up bunched posture by their classification. On the other hand, Mielke et al. (2016) employed a binary classification but considered the angle of the tongue blade as the primary feature with the categories tip/blade up /r/ and tip/blade down /r/, therefore avoiding the bunched

versus retroflex distinction entirely. Mielke et al. (2016) did note, however, that the blade raised configurations are the most ambiguous and in some instances, they used tongue-dorsum concavity, which is often present in tip down /r/, as a secondary indicator of bunching. Tongue-dorsum concavity or *sulcalization* (Catford, 2001) was also observed in Delattre and Freeman (1968) in bunched shapes, particularly in Type 4, labelled 'dorsal bunched with dip', which occurs between the pharyngeal and palatal constrictions. This 'dip' may also be observed in Type 5, a bunched configuration with a constriction produced with the tongue blade. We note that not all bunched shapes are articulated with sulcalization, an example of which is the 'dorsal bunched' Type 3 (as presented in Figure 2.2), which incidentally was the most common tongue shape in American English speakers reported in Delattre and Freeman (1968).

An alternative classification of tongue shapes was presented by Lawson et al. (2013) and included four categories, based on UTI data from Scottish English postvocalic /r/. They presented four configurations with two retroflex and two bunched shapes and were described as follows:

> Tip Up: the overall shape of the tongue surface is either straight and steep, or a concave shape, suggesting retroflexion.
>
> Front Up: the tongue surface forms a smooth convex curve. There is no distinct bunching of the tongue front or dip behind the front region.
>
> Front Bunched: the front of the tongue has a distinctly bunched configuration (the tip and the blade remain lower than the rest of the tongue front). A dip in the tongue's surface behind the bunched section is also apparent.
>
> Mid Bunched: the front, blade and tip are low, while the middle of the tongue is raised towards the hard palate. (Lawson et al., 2011, pp. 259-260)

Like other articulatory analyses, Lawson et al. (2011) also indicated that their four categories are on a continuum with Mid Bunched and Tip Up at the endpoints. In their first two bunched categories, the front to the mid-dorsum of the tongue may form the primary constriction for /r/,

while in the latter two, it is the tongue tip that forms the primary constriction. The description of the Front Up configuration seems most similar to Delattre and Freeman's Type 6 involving blade raising and, like Delattre and Freeman (1968), is considered retroflex by Lawson et al. (2011). In a later article by the same authors (Lawson et al., 2013), ultrasound tongue images depicting typical examples of the four categories were provided, which we present in Figure 2.3. The white tracing above the tongue contour represents the palate and the tongue tip is on the right. Interestingly, no distinction is made between 'real' sublaminal retroflexes and apical retroflexes, which do not present curling back of the tongue tip. They are instead grouped together under Tip Up and correspond to Delattre and Freeman's Types 7 and 8. The Front Bunched category with its 'dip' in the tongue's surface corresponds to Types 4 and 5, while the Mid Bunched category, with a lowered tongue tip, front and blade, without a 'dip' is probably closest to Type 3. Given the fact that ultrasound images show less articulatory information than cineradiographic ones – for example, hard structures such as the palate are not visible – this four-way classification is arguably better suited to UTI data. 8-way distinctions such as those presented by Delattre and Freeman (1968) would be challenging, if not impossible, to accurately classify based on data from UTI alone.

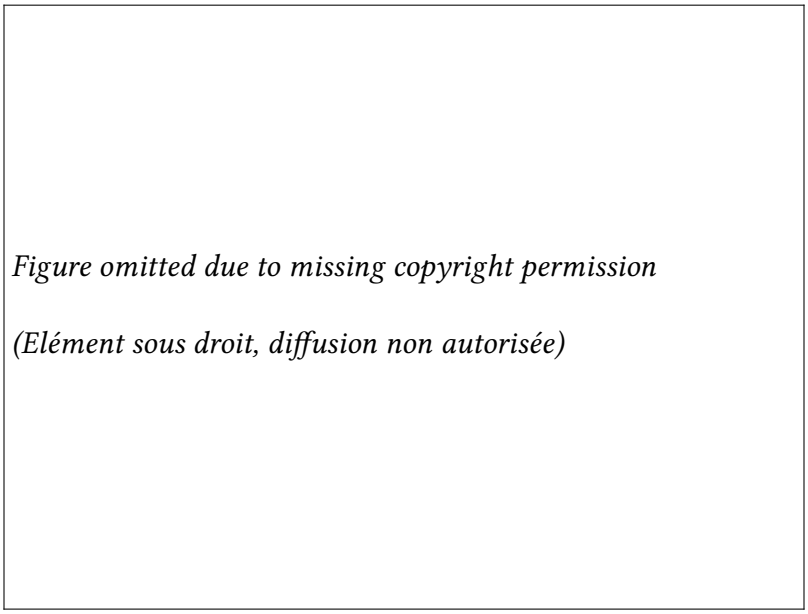



**Figure 2.3:** *Typical examples of tongue configurations for postvocalic /ɹ/ in Scottish English divided into four categories (from Lawson et al., 2013, p. 200).*

### 2.4.1 Factors constraining tongue shape

The numerous articulatory studies on English /r/ have shown that while some speakers use one tongue configuration exclusively in all contexts, others present consistent but individual variation conditioned by context, sociolinguistic factors, and perhaps even by physiology. Although certain speakers present similar patterns, /r/-allophony seems to be speaker specific, predominantly motivated by phonetic factors internal to the speaker (as discussed in Mielke et al., 2016). Magloughlin (2016) indicated that during acquisition, if a child's dominant strategy for /r/ proves ineffective, they may explore alternative articulatory configurations for those contexts, which become stable over time. Mielke et al. (2016) suggested that as bunched tongue shapes are favoured in environments which are the least compatible with retroflexion, retroflexion is the default articulation. Non-rhotic speakers of English present more retroflex shapes than rhotic speakers (e.g., Delattre & Freeman, 1968). As non-rhotic English produces /r/ in fewer contexts than in rhotic English, we can assume that non-rhotic English /r/ presents fewer contexts in which retroflexion may be incompatible. Furthermore, pre-vocalic /r/ is produced with higher retroflex rates than post-vocalic /r/. It is therefore possible that children

acquiring non-rhotic English are presented with fewer instances in which retroflexion is incompatible, making the use of retroflexion as their dominant strategy more likely than in children acquiring rhotic English, as suggested by Heyne et al. (2018) based on data from New Zealand English.

### Syllable position

As previously discussed, Delattre and Freeman (1968) and Hagiwara (1995) observed a general tendency for more retroflexion prevocalically, especially in word-initial position. Table 2.1 presents the percentage distribution of tongue shapes according to country of origin (England and the USA) and context (word-initial and word-final) based on the results presented in Delattre and Freeman (1968). The articulations deemed 'r'-less in Table 2.1 correspond to Types 1 and 2 which only occur in non-rhotic dialects, bunched articulations correspond to Types 3-5 and Types 6-8 are retroflex according to Delattre and Freeman's taxonomy. As Table 2.1 indicated, there is a strong tendency for retroflexion in word-initial position in both English and American speakers. This pattern is in stark contrast with word-final position, where /r/ is generally produced with a bunched tongue shape or not produced at all. Although American English speakers present more bunched shapes than retroflex ones in general, even bunchers show higher rates of retroflexion prevocalically with the use of 'fronted bunched' configurations (Delattre & Freeman, 1968). The same pattern was observed by Uldall (1958). She found her own 'molar' /r/ to occur postvocalically, while her 'tongue tip' /r/ occurred prevocalically. Other studies have also found a tendency for speakers to have retroflex /r/ in onset and bunched /r/ in codas (e.g., Mielke et al., 2016; Scobbie, Lawson, Nakai, Cleland, & Stuart-Smith, 2015; Westbury et al., 1998).

Mielke et al. (2016) suggested that the preference for retroflexion in syllable onsets may be motivated by the preference for larger (more constricted) anterior gestures in onset position, a phenomenon particularly prevalent in articulations involving multiple gestures such as nasals and laterals (Browman & Goldstein, 1995). An alternative but not contradictory explanation for increased retroflexion in onsets involves the process of syllable-initial augmentation. 'Artic-

| Context | Country | 'r'-less | Bunched | Retroflex |
|---|---|---|---|---|
| Word-initial | England | 0.00 | 0.00 | **100** |
| | USA | 4.00 | 21.06 | **74.93** |
| Word-final | England | **100** | 0.00 | 0.00 |
| | USA | 13.32 | **84.64** | 1.84 |

**Table 2.1:** *Percentage distributions of tongue shapes by context and country based on data presented in Delattre and Freeman (1968).*

ulatory strengthening', characterised by more extreme (i.e., less reduced) lingual articulations in consonants, has been shown to occur at the edges of prosodic domains, particularly word-initially in stressed syllables (Fougeron & Keating, 1997). A classic example of positional effects on gestural magnitude involves English /l/, in which clear /l/ as in *leap* occurs in onsets, while dark [ɫ] as in *peel* occurs in codas. Generative phonological accounts have considered the allophones of /l/ as two distinct phonetic entities, distinguishing the dark allophone from the clear with the features [+back] and [+high] (e.g. Chomsky & Halle, 1968). However, from articulatory data, Sproat and Fujimura (1993) found that a back lingual gesture is also present in clear /l/, as well as dark /l/. They argued that one of the features that makes clear and dark /l/ differ is not the presence of the articulatory gestures per se, but the magnitude of the gestures relative to one another. Sproat and Fujimura (1993) observed that in onset position, the more anterior coronal gesture was produced with greater magnitude than the more posterior dorsal one. In coda position, the opposite pattern was observed: the coronal gesture showed a reduction in magnitude with respect to the dorsal one. Indeed, in some speakers, the tongue tip gesture may be reduced or deleted entirely in coda, which sometimes results in complete vocalisation of /l/ (S. Giles & Moll, 1975; Lin & Demuth, 2013; Sproat & Fujimura, 1993; Wrench & Scobbie, 2003). In terms of magnitude, there is therefore a front-to-back pattern in onset and a back-to-front pattern in coda /l/.

The same pattern has been observed for English /ɹ/. Campbell, Gick, Wilson, and Vatikiotis-Bateson (2010) examined articulatory data from nine Canadian English speakers and found that in word-initial /ɹ/, the two most anterior gestures, i.e., the labial and dorsal (post-alveolar) gesture, had greater magnitude than the posterior pharyngeal one. In coda, the anterior gestures

were reduced in magnitude, while the pharyngeal gesture was strengthened. This pattern involving place and degree of articulation would therefore predict more extreme lip rounding for /r/ in onset than in coda position, which has indeed been widely observed (e.g., Delattre & Freeman, 1968; Lehiste, 1962; Proctor et al., 2019; Zawadzki & Kuehn, 1980).

In addition to the relationship between gestural magnitude and the syllable, a link between timing and magnitude has been proposed. For the allophones of /l/, Sproat and Fujimura (1993) found the coronal and dorsal gestures to occur almost simultaneously in onset position. On the other hand, the coronal gesture was shown to follow the dorsal one in coda. A much more recent real-time MRI study found that the coronal gesture precedes the dorsal one in onset, while the sequence is reversed in coda /l/ (Proctor et al., 2019). Browman and Goldstein (1995) suggested a 'general positional effect', in which the anterior tongue tip gesture for /t/, /n/ and /l/ in English is reduced in syllable-final position. Based on these observations, they indicated that there is a 'single syllable-final organisational pattern in which the wider constrictions always precede the narrower constrictions' (Browman & Goldstein, 1995, p. 167), explicitly linking intergestural timing and gestural magnitude in syllable coda.

Campbell et al. (2010) took this proposal one step further and found the same interaction to occur at all levels of the syllable in English /r/. The timing of articulatory gestures was observed to proceed sequentially from front-to-back in onset with the back constriction presenting gestural reduction. In coda, the two most anterior gestures exhibited reduced magnitude and the pharyngeal and labial gestures were produced before the dorsal one. Proctor et al. (2019) observed a slightly different back-to-front timing pattern for coda /r/: the pharyngeal gesture preceded the dorsal one (while there was no evidence of a labial gesture). Campbell et al. (2010) concluded that constriction width predicts gestural timing in English /r/: gestures with the greatest magnitude occur first. This proposal also accounts for the link between timing and magnitude of the articulatory gestures in English /l/. A simplified summary of the temporal and spatial findings involving English /r/ and /l/ reported in the literature is presented in Table 2.2.

| | Magnitude | Timing |
|---|---|---|
| Onset | front > back | front > back |
| Coda | back > front | back > front |

**Table 2.2:** *Simplified summary of temporal and spatial patterns in English* /ɹ/ *and* /l/ *reported in previous studies.*

### Segmental context

The segments which surround /ɹ/ may also have an influence on tongue shape due to coarticulation and articulatory ease. Delattre and Freeman (1968) reported higher retroflexion rates next to labial consonants, and the lowest next to coronals, followed by velars. Westbury et al. (1998) considered prevocalic /ɹ/ in point-tracking data from the x-ray Microbeam Speech Production Database (Westbury, Turner, & Dembowski, 1994) in 53 American English speakers. Less retroflexion and less extreme bunching were observed in the coronal cluster /stɹ/ than in word-initial /ɹ/. They concluded that when /ɹ/ is word-initial, the tongue is freer to move and able to reach more extreme articulations than when preceded by a consonant, but labial consonants are less constraining than coronals. Westbury et al. (1998) argued that as coronals place more constraints on the tongue tip position and angle than non-coronal consonants, retroflexion is less likely. Mielke et al. (2016) found retroflexion rates to be higher when /ɹ/ is not in a consonant cluster. When /ɹ/ does appear in a cluster, they noted that retroflexion is most frequent with labials than lingual consonants, especially in the context of coronals, where bunching is more likely. Indeed, Gick (1999) indicated that /ʃ/ and tip down /ɹ/ are both produced with tongue blade raising and lip rounding, which may facilitate bunching for /ɹ/ in the context of /ʃ/ (as discussed in Magloughlin, 2016). Likewise, according to Mielke et al. (2016), bunching often occurs next to segments produced with similarly bunched shapes, such as /ʃ/, /k/ and the vowel /i/.

Neighbouring vowels have also been shown to constrain tongue shape. Ong and Stone (1998) used UTI on one American English speaker to assess the influence of the following vowel. They observed bunching when /ɹ/ was flanked by front vowels /i ɪ ɛ æ/ and retroflexion

when flanked by back vowels /u ʊ o ɔ ʌ ɒ/. Mielke et al. (2016) found that retroflexion is favoured by open-back vowels and that the most natural contexts for retroflexion are /#rɑ/, /prɑ/, and other prevocalic contexts without lingual consonants. The authors noted that the most natural contexts for bunching are /ʃri/ and other postcoronal pre-/i/ contexts, along with most postvocalic contexts. They predicted that an American English speaker who retroflexes in the word *shriek* will retroflex in most if not all contexts, and a speaker who bunches in *rock* will bunch everywhere. The fact that retroflexion has been found to be incompatible with close-front vowels is perhaps not surprising as it has been suggested that retroflex sounds are always produced with a retracted tongue body (Hamann, 2002). Vowels which are also produced with a retracted tongue body such as back vowels, would therefore be more compatible. However, bunched /ɹ/ has also been associated with a retraction of the tongue. For example, Delattre and Freeman (1968) discuss the narrowing of the vocal tract in the pharyngeal region and much more recently, a retraction of the tongue body towards the lower rear pharyngeal wall was observed in all word-initial rhotics in a MRI study of four native American English speakers (Proctor et al., 2019). As both retroflex and bunched configurations are retracted, retraction cannot be the only articulatory property which makes retroflexion incompatible with front vowels. As Hamann (2003) suggested, the tongue shape for /iː/, which involves the tip being tucked under the lower front teeth, is inherently incompatible with that of retroflexion. Unlike in retroflexes, the tongue tip remains relatively low in the mouth for bunched /ɹ/, which is perhaps why bunching is more compatible with close-front vowels than retroflexion.

### Sociolinguistic factors

The covert nature of allophonic patterns for /ɹ/ makes the emergence of any dialectal patterns arguably unlikely. Indeed, as Mielke et al. (2016) pointed out, the fact that the difference between /ɹ/ allophones is not perceptible prevents convergence across speakers. As far as we are aware, no consistent dialectal patterns have been observed in American English. Westbury et al. (1998) had 28 speakers from Wisconsin in their dataset, who presumably spoke the same dialect, and yet no patterns regarding tongue shape emerged. Twist, Baker, Mielke, and Archangeli (2007)

conducted a perception study on American English speakers and found that although subjects can sometimes distinguish prevocalic from postvocalic /r/, they cannot distinguish a retroflex from a bunched /r/.

However, it has been observed that tongue shapes are socially stratified in the English spoken in the central belt of Scotland, which calls into question the covert nature of the bunched-retroflex dichotomy. In two UTI studies, Lawson et al. (2011) and Stuart-Smith, Lawson, and Scobbie (2014) observed that middle-class speakers used bunched articulations, while working-class speakers used more retroflex ones, and as a result, Lawson et al. (2011) argued that this articulatory variation must be in some way perceptible and exploited by listeners to index socio-economic class. Indeed, in a small-scale study, Lawson, Stuart-Smith, and Scobbie (2014) asked listeners to mimic speakers from audio recordings of middle- and working-class speakers and in some cases, mimicry participants adapted their tongue shape for /r/. Two of the authors in Lawson et al. (2011) classified the saliency of post-vocalic /r/ productions on a scale from weak to strong using the audio signal. They found a socially-stratified continuum of weaker working-class to stronger middle-class auditory variants of post-vocalic /r/. The use of opposing tongue shapes may be the cause of the perceived auditory difference. Lawson, Stuart-Smith, Scobbie, and Nakai (2018) and Lawson et al. (2013) noted that in addition to tongue shape variation, there is also temporal gestural variation between working- and middle-class speakers. In auditory weak /r/ tokens typical of working-class speakers who use retroflex shapes, the tongue-tip raising gesture is present but delayed with regards to voicing. Maximum displacement of the tongue tip for /r/ may occur after the offset of voicing and is therefore inaudible. In contrast, the maximum of the postvocalic /r/ gesture in middle-class speech occurs at or before the offset of voicing and, in some cases, occurs very early in the syllable rime. Just as /l/ may be vocalised in syllable coda due to the reduction of the tongue tip gesture, the acoustic saliency of /r/ may decrease when the maximum of post-alveolar gesture occurs later (Lawson et al., 2013; Lawson, Stuart-Smith, & Scobbie, 2018). In Lawson et al. (2013), the authors found that in middle-class speakers, the bunched tongue shape exerts a strong coarticulatory influence over preceding checked vowels. The location

and shape of the tongue during the onset of middle-class vowels closely resembles the location and shape of the following bunched /r/. The opposite is true for working-class speakers whose vowels do not share similar articulatory properties with the following retroflex tongue shape. It may be that children acquiring this variety of English are able to pick up on these temporal and gestural differences, resulting in stratification.

The only consistent dialectal pattern Delattre and Freeman (1968) reported is the one involving the three Anglo-English speakers who all produced the same tongue configuration in prevocalic /r/, Type 8, a sublaminal retroflex without phayngealisation. This 'extreme' retroflex tongue shape was not observed in the 43 American English speakers. The fact that there is a greater tendency for prevocalic /r/ to be retroflex than bunched, as previously discussed, may account for why non-rhotic speakers produce the highest rates of retroflexion. In the American participants presented in Delattre and Freeman (1968), speakers from the South exhibited the highest rates of retroflexion word-initially with 80 % using the tip up tongue shape (Type 7). In turn, they also had the highest degree of non-rhoticity word-finally among all the American English participants presented (37.13 %). As a result, Delattre and Freeman (1968) indicated that the South 'has the closest relationship with England with respect to /r/' (p. 62). Furthermore, in a recent large-scale UTI study of 62 New Zealand English speakers, nearly 20% of subjects produced exclusively retroflex tongue shapes (Heyne et al., 2018), a much higher proportion than the less than 8% exclusively retroflex American English users reported in a similar UTI study (Mielke et al., 2016). Heyne et al. (2018) speculated that as New Zealand English speakers very rarely produce /r/ in postvocalic environments, where bunching is heavily favoured, speakers are less likely to acquire bunched /r/ as an alternative articulation strategy if they have already mastered retroflexion.

### Physiological factors

A relationship between the shape of the hard palate and articulatory variability has been observed for certain speech sounds. A typical example involves the /s/-/ʃ/ contrast. Weirich and Fuchs (2013) observed that similar palatal morphologies such as those in monozygotic twins

yield similar articulatory realisations of the /s/-/ʃ/ contrast in German and that articulation is influenced by palatal steepness. As a result of these findings, some researchers have considered the impact of anatomical differences on English /r/ variability. Bakst (2016) considered the relationship between the curvature of the palate and the shape of the tongue for /r/ using a combination of dental casts and UTI. Curvature of the palate was not a significant predictor of retroflexion or bunching. However, speakers with flatter palates exhibited more consistent tongue shapes than speakers with more domed-shaped palates, although a domed palate did not necessarily predict increased variability. The articulation of /r/ in 80 native and non-native English speakers was recorded using MRI and the results are presented in Dediu and Moisik (2019). The authors suggested that anatomical aspects of the anterior vocal tract may influence articulation, particularly hard palate width and height, the overall size of the mouth, and the size of the alveolar ridge. However, their findings are tentative and require further verification in more native speakers. Indeed, other evidence suggests that physiology may not necessarily play a role. Magloughlin (2016) found that identical twins, who presumably have very similar, if not identical vocal tracts, may adopt opposing articulations in acquiring the sound. With the advent of real-time MRI, we can expect more important research of this kind to be undertaken in the future.

#### Summary of factors constraining tongue shape

As Mielke et al. (2016) pointed out, given the fact that bunched tongue shapes are favoured in environments which are 'the least articulatorily compatible with retroflexion' (p. 117), the retroflex [ɻ] may be considered the default allophone of English /r/. In North American English, deviations from this default (i.e., through the use of more bunched shapes) are generally motivated by phonetic factors internal to the speaker, particularly in relation to articulatory ease (Mielke et al., 2016). For example, retroflexion rate decreases when the tongue tip is constrained by coarticulation with neighbouring segments, such as with close-front vowels and coronal consonants. In strong prosodic contexts, such as the onset of stressed syllables in word-initial position, there is a universal tendency for more extreme and constricted anterior gestures,

i.e., 'articulatory strengthening', which perhaps motivates the use of increased retroflexion in these contexts. However, it should be stressed that although similar allophonic patterns emerge in some North American speakers, others have been shown to present one unique tongue shape regardless of context and in Scottish English, tongue shape has been found to be socially-stratified. However, tongue shape may also be idiosyncratic in nature, i.e., internal to the speaker, and may be due to individual patterns developed during childhood. As Magloughlin (2016) indicated, during acquisition, if a child's dominant strategy for /r/ proves ineffective, they may explore alternative articulatory configurations for those contexts, which become stable over time, resulting in individual allophonic patterns. Experience with /r/ in different contexts during acquisition may also account for the difference in retroflexion rate between non-rhotic and rhotic English speakers. It is possible that children acquiring non-rhotic English are presented with fewer instances in which retroflexion is incompatible, making the use of retroflexion as their dominant strategy more likely than in children acquiring rhotic English, as suggested by Heyne et al. (2018). Although more evidence is required, vocal tract morphology may also play a role in which tongue shape best allows a speaker to attain the acoustic output of a typical adult /r/. We may postulate therefore that in children, tongue shapes for /r/ may adapt with the changing size of the vocal tract until it reaches adult size.

Adaptive behaviour has been observed in adults when their habitual articulatory strategy for /r/ is mechanically perturbed. For example, Tiede, Boyce, Espy-Wilson, and Gracco (2010) fitted American English speakers with a palatal prosthesis and found that the majority of subjects responded by alternating between tongue shapes. The resulting formant values did not significantly differ from their unperturbed productions of /r/. The authors suggested that speakers acquire alternative production strategies during the exploratory period associated with childhood, which may remain 'in storage' and available for use when required, i.e., when speech is perturbed. However, Tiede et al. (2010) only considered the nonsense words 'ara' (/ara/), 'iri' (/iri/) and 'ooroo' (/uru/). As a result, we do not know if their subjects habitually use alternative tongue shapes in other phonetic contexts. The suggestion that articulatory strategies tried out in childhood remain in storage is therefore still to be determined.

## 2.5 Acquisition of /r/ in children

Very few articulatory accounts of the acquisition of /r/ in children currently exist. However, Magloughlin (2016) recorded four American children aged 3-6 years using UTI, who presented similar lingual variability to adults. Three out of four participants acquired postvocalic /r/ prior to prevocalic /r/. Similarly, acoustic data from a longitudinal study on nine young children acquiring American English suggest that the children progress towards postvocalic /r/ more rapidly than prevocalic /r/ (McGowan, Nittrouer, & Manning, 2004). It seems then that prevocalic /r/, which is the most frequent retroflexing context in adults, may be particularly challenging to acquire. Magloughlin (2016) concluded that if a child's dominant strategy for reaching adult-like targets proves ineffective in certain contexts, they may begin to explore alternative strategies, which become stable over time. One participant developed a secondary bunched configuration in contexts in which his dominant retroflex strategy proved ineffective, while another extended her habitual bunched articulation to all other contexts where she had a production lag. McGowan et al. (2004) speculated that the disparity in the acquisition of prevocalic and postvocalic /r/ may be due to limitations in motor control and in the morphology of the speech organs in young children. As Delattre and Freeman (1968) first observed, prevocalic /r/ tends to be more fronted than post-vocalic /r/, even in habitual bunchers. It is possible that the prevocalic context presents a challenge to children due to the dominance of the front of the tongue, which requires some time to mature, as noted by McGowan et al. (2004). Furthermore, the magnitude and timing of the articulatory gestures differ in prevocalic and postvocalic /r/, as discussed in Section 2.4.1 (p. 50). Speculatively, it may be that front-to-back temporal and spatial patterns are harder to acquire than back-to-front ones, although more articulatory evidence is required to corroborate this proposal. However, various studies have indicated that unlike /r/, English-speaking children acquire adult-like /l/ articulations in onset position before they produce adult-like coda /l/s (e.g., Dodd, Holm, Hua, & Crosbie, 2003; Dyson, Alice Tanner, 1988; Lin & Demuth, 2015; Smit, Hand, Freilinger, Bernthal, & Bird, 1990).

Like /l/, English /r/ is a well-cited example of a sound which is acquired late in children

(Boyce & Espy-Wilson, 1997). Indeed, Shriberg (1993) found that many children acquiring English do not reach adult-like productions of /r/ until age 8;0. Similarly, in their study on 3-9 year old American children from Iowa and Nebraska, Smit et al. (1990) reported that 90% of the children had attained correct /r/ production by 8;0. Other accounts place the age of acquisition of /r/ at around 6;0 (e.g., Vihman, 1996). Indeed, /r/ is commonly described as one of the most misarticulated sounds in children acquiring English as their first language (Boyce, Hamilton, & Rivera-Campos, 2016; Cialdella et al., 2020; Smit et al., 1990).

It is often reported that children substitute word-initial /r/ with [w] (Smit, 1993; Smit et al., 1990), at least according to perceptual judgements made by adults. However, acoustic and articulatory studies, although lacking, have indicated that children do differentiate between /w/ and /r/ in their production (Dalston, 1975; Kuehn & Tomblin, 1977), suggesting that perceptual judgements made by adults may not accurately reflect the phonetic realisation of /r/ in children. Indeed, Klein, Grigos, Byun, and Davidson (2012) showed that experienced and inexperienced clinicians differ in their perception of /r/ productions in American English speaking children. Disagreements were particularly apparent in their respective judgements of misarticulated productions of /r/. Many of the tokens rated as severely distorted or 'non-rhotic' productions of /r/ by experienced clinicians were deemed more acceptable (i.e., just 'distorted') by inexperienced listeners. The authors suggested that as many of these tokens are not obvious substitutions of another identifiable phoneme (i.e., /w/), the inexperienced listener will not necessarily consider them to be entirely misarticulated. Furthermore, adult-perceived /w/ substitutions in children's /r/ productions may actually be labiodental rather than labio-velar ones. Knight, Dalcher, and Jones (2007) present acoustic data from one speaker of SSBE between the ages of 3;8 and 3;11 and found that progress towards adult-like apical approximant /r/ is manifested through a gradual raising of F2 and a lowering of F3. This steady mastery of the acoustics of /r/ was also observed in American English speaking children in S. Lee, Potamianos, and Narayanan (1999) and McGowan et al. (2004). In Knight et al. (2007), the SSBE speaking child's /r/ development notably involved the elimination of [w] substitutions with concomitant increased labiodental realisations, and a decrease of F3-F2 distance. Their data suggested that

'developing speakers move gradually away from [w]-like articulations of /r/ to more adult-like articulations, producing a labiodental variant along the way' (Knight et al., 2007, p. 1581).

## 2.6 The pharyngeal component

Given the difficulty in visualising and measuring the pharynx, it is unsurprising that much less is known about the pharyngeal constriction for /r/ than the palatal one. However, it is generally agreed that in American English, a narrowing of the vocal tract in the pharyngeal region is involved in the articulation of /r/ (Boyce et al., 2016). For example, a pharyngeal constriction was observed in all the tongue configurations occurring in American subjects presented in Delattre and Freeman (1968). Later work has indicated that like the palatal constriction, the realisation of the pharyngeal constriction is variable. Alwan et al. (1997) found that there may be an interplay between the relative locations of the palatal and the pharyngeal constriction. American English speakers who produce the palatal constriction with the posterior tongue body had a more inferior pharyngeal constriction than those in which the palatal constriction occurs further front. It was also noted by both Delattre and Freeman (1968) and Alwan et al. (1997) that tongue shapes with large degrees of sulcalization tend to have a higher pharyngeal constriction than those without. Furthermore, both of these studies found that extreme sublaminal retroflexes (i.e., the one associated with Anglo-English speakers in Delattre and Freeman (1968)) may not present a pharyngeal constriction at all.

As discussed in Section 2.4.1 (p. 50), the timing and magnitude of the pharyngeal gesture in relation to the lingual and labial ones associated with /r/ has been assessed (e.g., Campbell et al., 2010; Proctor et al., 2019). Temporal and spatial patterns have been linked to the position of /r/ in the syllable. The pharyngeal constriction, occurring after the labial and lingual components, is reduced in syllable onset. In contrast, in coda, the pharyngeal gesture is produced with greater gestural magnitude than the lingual and labial ones, and generally occurs before the other two gestures.

A link may be made between the magnitude of the pharyngeal gesture in syllable coda and

vocalisation of /r/, or r-loss, in this post-vocalic context, which historically occurred in non-rhotic Englishes. When vocalisation of /r/ takes place following a non-open vowel, the historical /r/ is replaced with schwa, e.g., *here* *hiɹ >[hiə] (Gick, 2002b). It has been hypothesised by McMahon, Foulkes, and Tollfree (1994), and later by Gick (1999), that if the palatal constriction associated with /r/ were removed, the remaining tongue configuration would closely resemble the articulation of schwa (Gick, 2002b). Indeed, Gick, Kang, and Whalen (2000) and Gick (2002b) present articulatory evidence which supports the existence of an /r/-like pharyngeal constriction in the articulation of schwa in American English. The 'general positional effect' proposed by Browman and Goldstein (1995) may therefore account for vocalisation of /r/, in which anterior articulatory gestures are reduced in syllable-final position, leaving the posterior, pharyngeal one for /r/, which closely resembles the articulation of a schwa.

As it has been claimed that Anglo-English /r/ may not involve a pharyngeal component (Delattre & Freeman, 1968), we do not intend to report on the pharynx in this thesis. Although as far as we are aware, this suggestion has yet to be replicated elsewhere, at the very least, the pharyngeal constriction is considered to be reduced in pre-vocalic /r/, the only context in which /r/ is produced in Anglo-English. We can therefore assume that the pharyngeal constriction is reduced relative to the labial and lingual ones in Anglo-English /r/. Furthermore, while real-time MRI would allow us to image the pharynx, no such data currently exists for Anglo-English /r/. This thesis will therefore focus on the palatal and labial constrictions, which we intend to observe via UTI and lip camera data.

## 2.7 THE LABIAL COMPONENT

Although the vast majority of articulatory work on /r/ focuses on its lingual gesture (Docherty & Foulkes, 2001), it is generally agreed that /r/ may be labialised but the exact phonetic implementation of labialisation is unknown. It has been observed that lip rounding is likely to occur in prevocalic and pre-stress syllable positions in both American English (Delattre & Freeman, 1968; Mielke et al., 2016; Proctor et al., 2019; Uldall, 1958; Zawadzki & Kuehn, 1980) and Anglo-English (Abercrombie, 1967; Jones, 1972; Scobbie, 2006), regardless of the shape

of the tongue. In Anglo-English, Scobbie (2006) informally observed that between 25 and 50 percent of nonbroadcasters interviewed on United Kingdom radio and television labialised /r/ at least some of the time. In contrast, Gimson (1980) suggested that lip rounding in Anglo-English /r/ is largely conditioned by the quality of the following vowel, with /r/ preceding rounded vowels exhibiting more rounding than /r/ preceding non-rounded vowels. However, it has been observed that English speakers do not always round their lips for so-called rounded vowels (Brown, 1981), and that they use less rounding than speakers of other languages with phonologically equivalent rounded vowels, such as French (Badin, Sawallis, & Lamalle, 2014; I. L. Wilson, 2006). Ladefoged and Disner (2012) noted that modern productions of the vowel /uː/ have relatively spread lips in comparison to productions of the recent past, although articulatory studies have indicated that while /uː/ remains rounded, it is no longer a back vowel (e.g., Harrington, Kleber, & Reubold, 2011; King & Ferragne, 2018; Lawson, Stuart-Smith, & Rodger, 2019). Brown (1981) even went as far as to suggest that the main origin of lip rounding in English derives not from rounded vowels, but rounded consonants, and that the most marked lip movement can be found in the consonants /ʃ, tʃ, ʒ, dʒ/ and /r/, although this idea does not seem to have been developed further.

English pronunciation manuals vary with their treatment of the labial gesture. O'Connor (1967) recommends learners approach [ɹ] from [w], and then curl the tip of the tongue back until it is pointing at the hard palate, which presumably supposes that the lip postures for [ɹ] from [w] are identical. Others warn learners not to exaggerate rounding for /r/ because it would have the effect of producing the percept of a [w] (e.g., Lilly & Viel, 1977; Roach, 1983). While Ehrlich and Avery (2013) indicate that lip rounding is a possibility, Ashton and Shepherd (2012) inform learners that using their lips to help them form the /r/ sound is 'wrong' and recommend learners use their fingers to hold their lips still in order to practise using just their tongue (p. 49).

The terms *lip protrusion* and *lip rounding* seem to be used interchangeably in descriptive accounts of English /r/, perhaps because, as Laver (1980) indicated, protrusion without lip rounding is rare in the world's languages. However, inspired by Sweet (1877)'s articulatory

account of rounding in vowels, Brown (1981) explicitly differentiated the two: rounding restricts lip aperture by compressing the lip corners, but does not necessarily push the lips forward, as is the case for English /w/; while protrusion pushes the lips forward, opening and everting them to show the soft inner surfaces, as in English /ʃ, tʃ, ʒ, dʒ/ and /r/. Again like Laver (1980), Brown (1981) essentially used horizontal compression to define lip rounding, which is notably absent from her description of the 'protruded' consonants /ʃ, tʃ, ʒ, dʒ/ and importantly for the present study, /r/. However, in a very recent articulatory study on sound change triggered by American English /r/, B. J. Smith, Mielke, Magloughlin, and Wilbanks (2019) observed that /ʃ/ lip rounding is different from /r/ lip rounding. Their speakers produced /ʃ/ with open protruded ('outrounded') lips, while /r/ involved vertical movement by the upper and/or lower lip, sometimes with a narrow lip aperture ('inrounded'). However, both /ʃ/ and /r/ exhibited inter-speaker variability in the shape and area of the labial constriction.

## 2.8 ACOUSTIC PROPERTIES

Despite the diversity of possible tongue shapes observed for post-alveolar /r/, the acoustic profile of these different tongue configurations is remarkably indistinguishable, at least with regards to the first three formants (Espy-Wilson et al., 2000). It is generally agreed that the most salient acoustic feature for /r/ is its low third formant (F3) value, usually below 2 000 Hz (Boyce & Espy-Wilson, 1997; Delattre & Freeman, 1968; Proctor et al., 2019) and some researchers have remarked on the close proximity of F3 to F2 (Dalston, 1975; Guenther et al., 1999; Lisker, 1957; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957; Stevens, 1998). An alternative account suggests that the percept of /r/ is defined not by F3, but by a single dominant peak in the F2 frequency region (Heselwood & Plug, 2011). Formant values from American English /r/ reported in the literature across tongue shapes, phonetic contexts and sexes range from 300-500 Hz for F1, 900-1 300 Hz for F2, and 1 300-2 000 Hz for F3 (Delattre & Freeman, 1968; Espy-Wilson, 1992; Espy-Wilson & Boyce, 1999; Uldall, 1958; Westbury et al., 1998; Zhou et al., 2008). In rhotic Englishes, prevocalic /r/ presents lower formant values than postvocalic /r/, which is generally assumed to be the result of the presence of lip rounding in prevocalic /r/

(Delattre & Freeman, 1968; Lehiste, 1962; Zawadzki & Kuehn, 1980). As far as we are aware, no study has observed systematic differences between retroflex and bunched /r/ up to the third formant. However, beyond F3, Espy-Wilson and Boyce (1994) found that F3 and F4 are further apart for retroflex than they are for bunched /r/. More recently, consistent acoustic differences have been found in the higher formants in American English. Notably, the difference between F4 and F5 has been found to be larger in retroflex than in bunched /r/. Zhou et al. (2008) found that retroflex /r/ in American English males showed a difference between F4 and F5 of over 1 400 Hz compared with 700 Hz for retroflex /r/. This result has since been replicated in studies on postvocalic /r/ in Scottish English (Lawson, Stuart-Smith, & Scobbie, 2018; Lennon, Smith, & Stuart-Smith, 2015).

A variety of attempts have been made to account for the acoustics of English /r/, particularly with regards to the maintenance of the low F3 values observed across a multitude of articulatory configurations. Accounts for the source of the low F3 associated with /r/ have been proposed using both Perturbation Theory (e.g., Johnson, 2012; Ohala, 1985) and multi-tube models (e.g. Alwan et al., 1997; Espy-Wilson et al., 2000; Stevens, 1998) with varying degrees of success. Perturbation Theory relates vocal tract constrictions to formant frequencies by accounting for perturbations to a uniform, unconstricted tube, where one end is closed and the other end is open (i.e., a quarter-wavelength resonator). Perturbation Theory states that if you constrict the tube at a place along its length where there is a point of maximum velocity (or zero pressure), i.e., at the location of an antinode, the frequency of the corresponding resonance will fall. Conversely, if you constrict a tube at a place along its length where there is a point of maximum pressure (or zero velocity), i.e., at the location of a node, the frequency of the corresponding resonance will rise (Chiba & Kajiyama, 1941). Perturbation Theory predicts the points of maximum velocity for F3 to occur in the pharyngeal, palatal and labial regions, which, according to Johnson (2012) 'nicely illustrates' the utility of Perturbation Theory in that a combination of all three constrictions are used for English /r/. Perturbation Theory would thus predict that the source of the low F3 typical of /r/ is a combination of all three constrictions, which is indicated by the distribution of antinodes for F3 in Figure 2.4. However,

Espy-Wilson et al. (2000) used area functions from MRI data to show that Perturbation Theory cannot adequately account for the actual constriction locations speakers use. For example, they found that the palatal constriction is actually located at a point of maximum pressure (i.e., at a node) and not maximum velocity (i.e., at an antinode), which, according to Perturbation Theory, would more likely raise F3 than lower it.

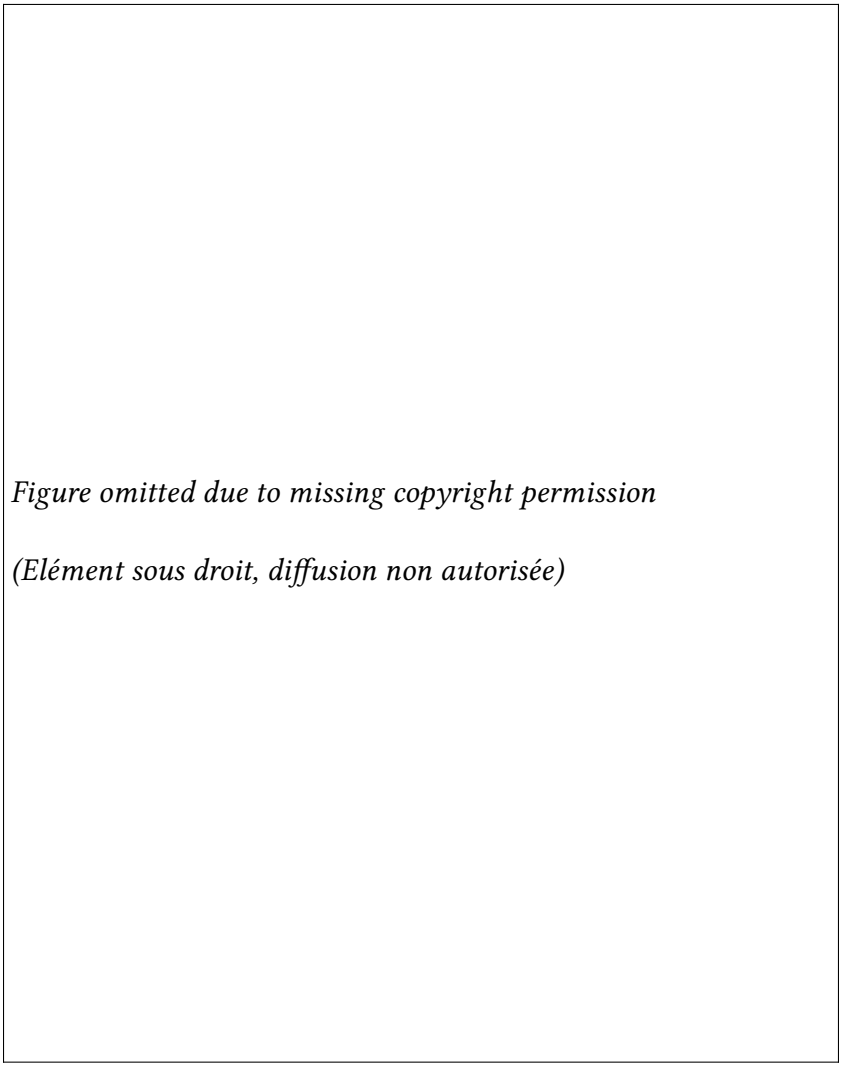



**Figure 2.4:** *Locations of nodes and antinodes in a tube open at one end in the unconstricted vocal tract. Perturbation Theory predicts that a constriction at the location of an anitnode (labelled* A*) in the vocal tract would lower the frequency of the corresponding resonances. Nodes are indicated by the intersections of the sine waves (adapted from Johnson, 2012, Figure 6.7).*

Contrary to Perturbation Theory, multi-tube models consider the vocal tract to comprise of several tubes of different areas and lengths, and that the source of the different formants is the resonating frequency of the different tubes (Espy-Wilson et al., 2000). Multi-tube model accounts have affiliated the low F3 typical of /r/ with the front cavity, i.e., between the palatal constriction and the lips. Stevens (1998) found that F3 results from a large front cavity volume for /r/, although he suggested that the various tongue configurations used for /r/ do not lower F3 per se, but introduce an extra resonance, $F_R$, in the frequency range normally occupied by F2 with a drop in amplitude of F3 proper. Based on speakers' actual vocal tract dimensions derived from MRI data, Espy-Wilson et al. (2000) developed a multi-tube model to account for cavity affiliations for /r/. With regards to F3, their model confirmed that F3 is indeed a front cavity resonance, which includes a lip constriction formed by the tapering gradient of the teeth and lips – with or without rounding – and a large volume cavity behind it that includes a sublingual space. They found that this sublingual space acts to increase the volume of the cavity and lowers F3 by approximately 200 Hz. Interestingly, while Perturbation Theory would predict that a constriction in the pharyngeal region would lower F3, Espy-Wilson et al. (2000)'s model indicates that eliminating the pharyngeal constriction has minimal effect on F3.

Physical models of the vocal tract have also indicated that the size of the front cavity has an influence on F3. Lindblom, Sundberg, Branderud, Djamshidpey, and Granqvist (2010) noted that despite the advances in articulatory-acoustic relations particularly as a result of work by Gunnar Fant, our understanding of vocal tract acoustics remains incomplete with respect to the treatment of lip spreading and of the sublingual space. As a result, they created a physical twin-tube model in order to model acoustics. Their results corroborate multi-tube models of /r/ in that they too associate the front cavity with F3. When the volume of the front cavity is manipulated, all the while maintaining the lip opening area at a constant (1 $cm^2$), the lowest F3 values are observed with the largest possible front cavity volumes. In essence, their physical model of the vocal tract shows that the sublingual space contributes to the overall area of the front cavity and that when the volume of the front cavity increases, F3 decreases. Interestingly, they observed an interaction between the size of the sublingual cavity and the degree of lip

spreading. The lowest possible F3 values occur with the lowest degree of spreading. However, the main acoustic correlate of spreading, according to their physical model, is F2: F2 increases as the lips become more spread.

The consistency in formant values observed for /r/ has given rise to the suggestion that trading relations may exist between the different articulatory manoeuvres which reciprocally contribute to the lowering of F3. Dependence on one of these articulatory manoeuvres would be accompanied by less of another, and vice versa (Tiede et al., 2010). In an acoustic and articulatory study of the production of /r/ in seven American English speakers, Guenther et al. (1999) observed systematic trade-offs between the length of the front cavity and the length and size of the constriction, which allowed speakers to maintain stable F3 values across different contexts of /r/. As a result, articulatory variability is juxtaposed with acoustic stability. Speakers modify the length of the front cavity and the length of the constriction in order to achieve the necessary total volume of the cavity which produces the low F3 typical of /r/ (Matthies et al., 2008). The results from Guenther et al. (1999) therefore suggest that the target of speech production is acoustic in nature, as opposed to the traditional view, which would consider each phoneme to have a canonical vocal tract shape target, as Guenther et al. (1999) discussed.

Tongue shapes with a raised tongue tip create a cavity underneath the tongue blade, the sublingual space. Since the reported tongue shapes for /r/ vary with respect to the elevation of the tongue tip, from tip down bunched to curled up retroflex, it is likely that the size of the sublingual space varies across tongue shapes. Extreme retroflex shapes with sublaminal articulations would presumably have a larger sublingual space than apical ones, as briefly discussed in Espy-Wilson et al. (2000). Similarly, unlike tip up /r/, the tongue tip is down in bunched /r/ and therefore has negligible sublingual space (Zhang, Boyce, Espy-Wilson, & Tiede, 2003). Indeed, Alwan et al. (1997) used MRI- and Electropalatography (EPG)-derived vocal tract dimensions, and in one American English speaker, the front cavity volume was larger for retroflex than bunched /r/ (6.1 $cm^3$ and 4.5 $cm^3$, respectively). This difference may be due to the smaller sublingual space in bunched /r/, although Alwan et al. (1997) did not explicitly make this suggestion. Trading relations involving the sublingual space may therefore

be possible. Given the impact of the presence of a sublingual space on F3, Alwan et al. (1997) posited a trading relation between the sublingual space for tip up /ɹ/ and a more posterior palatal constriction for tip down /ɹ/, which was also discussed by Espy-Wilson et al. (2000). Extending the front cavity – and thus increasing its volume – could also be achieved through the formation of a separate lip protrusion channel (Espy-Wilson et al., 2000). Yet, to the best of our knowledge, trading relations involving lip protrusion have yet to be investigated, which is perhaps due to the lack of available lip data, as Espy-Wilson et al. (2000) pointed out.

## 2.9 Labiodental variants

Labiodental articulations of /ɹ/, i.e., involving the lower lip and the upper teeth such as [ʋ], are predominantly associated with Anglo-English. The earliest known commentaries on labiodental-like variants in Anglo-English date back to the mid-1800s (for a diachronic review of labiodentalisation, see Foulkes & Docherty, 2000). Up until the early 2000s, labiodentalisation was dismissed as a speech defect or an infantilism – due to its presence as a development feature in children acquiring English (Kerswill, 1996; Knight et al., 2007) – or as an affectation of upper class speech (Foulkes & Docherty, 2000). For example, Jones (1956) treated the labiodental variant as 'defective' and suggested strategies for its correction (as cited in Armstrong & Pooley, 2013). However, Foulkes and Docherty (2000) presented evidence to suggest that not only are perceptions of labiodental /ɹ/ changing, particularly in the popular media, but [ʋ] is now a relatively widespread feature in non-standard south-eastern accents of England, which was also suggested by Wells (1982). Indeed, as Armstrong and Pooley (2013) noted, where the labiodental variant was once stigmatised as defective, it is now treated with greater tolerance to such an extent that 'many parents may now be less ready to correct this variant as defective in their children's speech' (p. 142).

Furthermore, Foulkes and Docherty (2000)'s review of dialectological studies indicated that labiodentalisation is spreading from its south-eastern epicentre to other urban accents across England. Instances of [ʋ] have been noted in several areas outside the capital including Milton Keynes, Reading, Hull (Williams & Kerswill, 1999), Norwich (Trudgill, 1974, 1988, 1999b),

Derby (Foulkes & Docherty, 2000), Leeds (Marsden, 2006), Middlesborough (Llamas, 1998) and Newcastle (Foulkes & Docherty, 2000). Foulkes and Docherty (2000) hypothesised that [ʋ] may be spreading as part of a general levelling process which is currently occurring in Anglo-English. This levelling process is believed to have originated in non-standard south-eastern varieties, which according to Foulkes and Docherty (2000), enjoys sociolinguistic dominance in young people across urban areas of England. This accent levelling typically affects consonants and the most famous features include TH-fronting, /l/-vocalisation and /t/-glottaling. As Foulkes and Docherty (2000) pointed out, /r/-labiodentalisation may well be part of this same general levelling process.

While sociolinguistic factors are no doubt at play, few phonetic accounts as to why labiodental variants are rapidly emerging currently exist. It is generally implied that labiodental variants have emerged by speakers retaining the labial component of /r/ at the expense of the lingual one (Docherty & Foulkes, 2001; Foulkes & Docherty, 2000; Jones, 1972), although articulatory data is lacking. This proposition would imply that the lip posture for /r/ is labiodental, i.e., produced with an approximation between the lower lip and the upper front teeth, regardless of whether or not there is an accompanying lingual gesture, which cannot currently be confirmed due to the lack of articulatory data. Docherty and Foulkes (2001) hypothesised that this change in progress from [ɹ] to [ʋ] may be the result of the heavy visual prominence of the labial gesture for /r/, which may have led to the labial taking precedence over the lingual articulation. Lindley and Lawson (2016) observed one participant who produced labiodental /r/ with no observable tongue body gesture. However, another participant presented labiodentalisation accompanied by a tip up tongue configuration, leading them to suspect that the change in progress from [ɹ] to [ʋ] may be phonetically gradient, in line with Docherty and Foulkes (2001)'s hypothesis.

Phonetic analyses of labiodental variants are few and far between and generally do not extend much beyond auditory accounts. Foulkes and Docherty (2000) and Marsden (2006) rated the perceptual quality of /r/ on a 4-point auditory scale according to the degree of alveolar or labial articulation. However, Foulkes and Docherty (2000) also included a spectrographic and formant analysis of labiodental Anglo-English /r/, which was probably the first study to do

so. They found that while energy in the higher frequencies beyond F3 is relatively weak for alveolar /ɹ/, in labiodental /ɹ/, high frequency energy is much clearer. Like in other studies, they categorised alveolar /ɹ/ as having a low F3 in close proximity to F2 (at around 1 700 Hz). Labiodental /ɹ/, in contrast, had a markedly higher F3, at around 2 200 Hz. They observed a clear correlation between their auditory index score and their acoustic measurements: variants which gave the auditory impression of [ʋ] had higher F3 values. Foulkes and Docherty (2000) argued that this result is expected as we would predict articulations lacking retroflexion or bunching of the tongue to result in higher F3 values.

Exposure to labiodental variants without a canonically low F3 may have resulted in a shift in the perceptual weighting of /ɹ/ in England. Somewhat unexpected differences have been observed in the perception of approximants between American English and Anglo-English listeners. In Dalcher et al. (2008), American and English participants judged whether copy-synthesised sounds with manually adjusted formant values were more like /ɹ/ or /w/. A significant difference was observed for a stimulus which had a third formant typical of /ɹ/ (1 682 Hz) and second formant typical of /w/ (725 Hz). American speakers identified this stimulus as /ɹ/ 90% of the time, while Anglo-English speakers only identified it as /ɹ/ 59% of the time. Dalcher et al. (2008) suggested that the reason for such a disparity may be due to the exposure to labiodental variants without a canonically low F3 in Anglo-English listeners. The increase in /ɹ/ variability with respect to its third formant may have served to catalyse a cue-shift from F3 to F2 in the perception of the /ɹ/-/w/ contrast. As Figure 2.5 suggests, apical productions of /ɹ/ contrast with /w/ both with respect to F2 and F3. However, F3 is no longer contrastive in labiodental productions. As a consequence, Dalcher et al. (2008) speculated that a low F3 alone is no longer a sufficient cue to distinguish /ɹ/ from /w/ in Anglo-English and that the F2 boundary between /ɹ/ and /w/ may have become sharper in Anglo-English speakers. As such, a token with a low, [w]-like F2 value would be perceived as /w/ by Anglo-English listeners even when accompanied by a low [ɹ]-like F3.

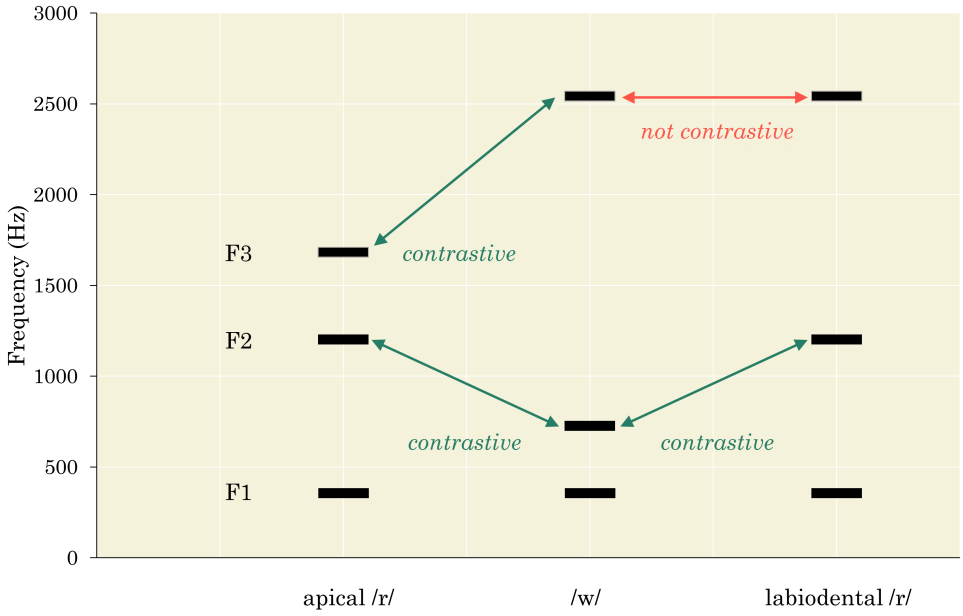



**Figure 2.5:** *Formant contrasts between* /r/ *and* /w/ *pronunciation variants based on the formant values presented in Dalcher et al. (2008).*

## 2.10 CHAPTER CONCLUSION

The review of existing work presented within this chapter has indicated that despite the vast array of existing phonetic studies, our understanding of English /r/ is still incomplete, particularly concerning its production in non-rhotic Englishes and the accompanying labial gesture. It has been well-documented in rhotic Englishes that the post-alveolar approximant may be produced with a variety of different tongue shapes from tip up retroflex to tip down bunched. Although tongue shape is generally thought to be speaker-specific (at least in American English), variation may be conditioned by coarticulation, syllable context, sociolinguistic factors and perhaps even by speaker physiology. Despite the diversity of possible tongue shapes, the acoustic profile of post-alveolar /r/ is remarkably consistent and is characterised by a particularly low third formant, usually in the frequency region which is normally occupied by F2. Articulatory-acoustic models have associated this low F3 with a large volume front cavity. The different possible tongue shapes for post-alveolar /r/ may result in differing sized

front cavities. To obtain a stable acoustic output across post-alveolar /r/ variants, speakers may make systematic trade-offs between the articulatory manoeuvres available to them which reciprocally contribute to the lowering of F3. Possible manoeuvres may include modifications to the size of the sublingual space, to the place and length of the lingual constriction and to the length of the lip protrusion channel. If a trading relation involving the lips and tongue exists, it is possible that systematic differences in lip protrusion may be observed across the possible tongue shapes associated with /r/, which has yet to be examined.

The lips seem to play an important role in the production and perception of /r/ in Anglo-English because labiodental variants are becoming increasingly common. Exposure to these variants may have had an effect on the perception of /r/, particularly with regards to the relative importance of F3 as an acoustic cue. It is thought that labiodental /r/ may have emerged due to the visual prominence of the lips in the 'standard' post-alveolar variant, although no detailed phonetic account of the lips for /r/ currently exists for any variety of English. As a result, the contribution of the lips to both the production and to the perception of /r/ in Anglo-English will be investigated in this thesis.

# Phonetic accounts of labialisation 3

As Westbury and Hashi (1997) remarked, 'the lips pucker and spread and rise and fall during speech' (p. 405) permitting speakers to alter the shape of the vocal tract and modify the acoustics of the sound they produce. The lips are also a visible articulator, contributing supplementary phonetic information and increasing the perceptibility of speech visually. Therefore, as Honda, Kurita, Kakita, and Maeda (1995) pointed out, the action of the lips for speech provides a useful means of investigating multimodal aspects of speech production and perception. Low-level phonetic descriptions of the lips have revealed that behind the apparent simplicity of the binary phonological feature [± round] lies a complex pattern of articulatory variability resulting from inter-speaker, contextual and cross-linguistic differences (Zerling, 1992). While lip rounding is more closely associated with vowels, the equivalent labial activity found in consonants is known under the term *labialisation*. Consonants are described as labialised when they are accompanied by a secondary labial gesture and as a result, are often transcribed phonetically with the diacritic [ʷ]. On the other hand, consonants described as labial are those whose primary articulation occurs at the lips, e.g., in bilabials such as [p], [b], [m] and in labiodentals such as [f], [v], [ʋ].

One of the main goals of this thesis is to assess the contribution of the accompanying labial

gesture to the production and perception of English /r/, an approximant consonant. As a result, in this chapter we focus our attention on labialisation as a secondary articulation, which is generally associated with consonants. However, as detailed phonetic accounts of labialisation in consonants are somewhat scarce, we will supplement this review with phonetic accounts of lip rounding in vowels. Indeed, the articulatory dimensions used for rounding in vowels are the same as those used for labialisation in consonants (e.g., Marchal, 2009). Furthermore, lip rounding in vowels can arguably be considered to be a secondary articulation as like labialised consonants, they too are coupled with a tongue body gesture. The review of previous studies will indicate that both the vowels and consonants traditionally described as 'rounded' and 'labialised' may not actually be produced with a 'rounded' lip configuration. As a result, we choose to employ *labialisation* as a general term for both vowels and consonants, which we consider to be a more phonetically neutral label. We will show that different lip configurations may have different acoustic consequences, which are dependent on interactions with the lingual constriction.

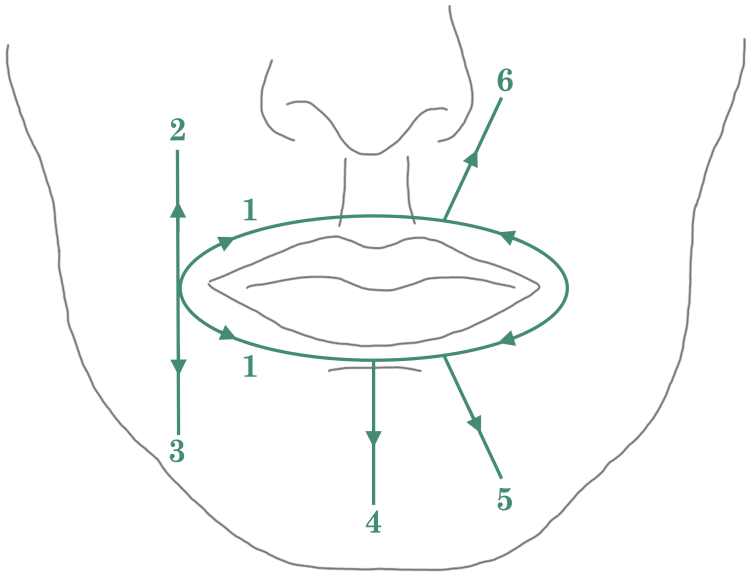
## 3.1 PRINCIPAL MUSCLES INVOLVED IN THE LIP MOVEMENTS FOR SPEECH

Before reviewing existing phonetic descriptions of labialisation, we will briefly describe the principal muscles involved. Figure 3.1 depicts the main muscles used in lip opening and closing based on descriptions in Honda et al. (1995) and Laver (1980). As detailed in their *Dissection Manual for Students of Speech*, Ladefoged, Epstein, and Hacopian (2002) noted that from a phonetic viewpoint, there are three major movements of the lips, which we summarise as follows:

1. *Rounding and spreading* (the lip corners are drawn together or pulled apart). Rounding is largely achieved by the orbicularis oris, which encircles the lips and acts as a sphincter. The orbicularis oris is described as the kissing and whistling muscle in Marieb and Hoehn (2007). The contraction of this muscle (i.e., for rounding) is associated with a pronounced wrinkling of the labial skin (Folkins, 1978).

2. *Protrusion* (the lips are pushed forward, extending the vocal tract). The lower lip is more implicated in protrusion than the upper. The action of the lower lip is predominantly controlled by the mentalis and the depressor labii inferioris muscles. According to Marieb and Hoehn (2007), protrusion of the lower lip with the mentalis muscle results in wrinkling of the skin of the chin.

3. *Vertical compression* (the lips come together, predominantly by the raising of the bottom lip). Vertical compression occurs without lip rounding mainly by raising the lower lip while raising the jaw. Some vertical compression can be achieved without jaw raising mainly by the actions of the inferior part of the orbicularis oris and the mentalis muscle.




1. orbicularis oris
2. levator anguli oris
3. depressor anguli oris
4. mentalis
5. depressor labii inferioris
6. levator labii superioris

**Figure 3.1:** *Schematisation of the principal muscles involved in lip opening and closing.*

## 3.2 MEASURING THE LIPS

Non-invasive tracking of muscle activity in the lips during speech has been made possible via surface Electromyography (EMG) in which the electric potential generated by muscle cells is detected via electrodes. However, modern EMG studies are rather rare, perhaps due to the difficulty in accurately placing electrodes.[1] As Cattelain, Garnier, Savariaux, Gerber, and Perrier (2018) noted, the anatomy of the face is particularly complex: many muscles overlap, particularly around the lips, and inter-speaker variability is high. As a result, an electrode may track the activity of a neighbouring muscle even if the electrode is not situated directly above it. Analysing the data is also not without difficulty, as labial muscle activity is composed of multiple sub-movements. O'Dwyer, Quinn, Guitar, Andrews, and Neilson (1981) were the first to provide anatomic criteria for the correct placement of electrodes, but as Cattelain et al. (2018) pointed out, inter-speaker variability requires speaker-specific adjustments to electrode placement. Despite these challenges, it has been shown that EMG conveys sufficient information to predict 3D lip shapes (Eskes et al., 2017), suggesting that it is a powerful technique which could be considered in future studies.

Other methods of investigation have been employed to capture and quantify the lip movements associated with speech. Given the fact that the lips are a visible articulator, it is not surprising that video recordings of the lips are a common imaging technique. 2D measurements can be made for horizontal and vertical lip aperture from still frontal lip camera images either manually (Mayr, 2010) or automatically, using contour detection and extraction techniques (Klause, Stone, & Birkholz, 2017). Similarly, lip protrusion can also be measured from profile lip camera images (Lawson et al., 2019; Saitoh & Konishi, 2010). Measurements may also be made of the lips and jaw using flesh-point techniques such as optical motion tracking, where the position of markers placed on a speaker's face and lips can be tracked. Noteworthy studies employing such a technique include Georgeton and Fougeron (2014), which examines the effect of prosody on lip rounding in French vowels, and Campbell et al. (2010) in which the labial gesture for North American English /r/ was measured. An alternative point tracking technique

[1] Our own attempts have proved inconclusive.

capable of measuring the position of the lips is Electromagnetic Articulography (EMA). Given the fact that EMA sensors can be placed on multiple articulators, e.g., the tongue tip, the tongue body and the lips, it is possible to measure their temporal coordination (Kochetov, 2020). However, as Noiray, Cathiard, Ménard, and Abry (2011) pointed out, lip shape and constriction cannot be adequately tracked with flesh-point measures alone. As a result, Noiray et al. (2011) supplemented flesh-point tracking with a video shape tracking system, in which the lips are painted in blue to maximise the colour contrast with the skin (Lallouache, 1991). In post-processing, the blue lip shapes are tracked to calculate lip aperture, interlabial area and lip protrusion (Noiray, Ries, & Tiede, 2015).

An ideal technique would naturally be one which allows us to capture and measure the entire vocal tract with a sufficiently high temporal and spatial resolution, which is currently technologically challenging, costly, and relatively invasive (Kochetov, 2020). Advances in real-time MRI technology may make entire vocal tract measures more of a possibility in the near future. Indeed, recent MRI studies have been undertaken which consider labial articulation including Proctor et al. (2019) on American English /r/. However, data collection and analysis is particularly challenging and as a result, sample sizes tend to be small.

## 3.3 The articulation of labialisation

The vast majority of existing phonetic descriptions of labialisation consider the lip rounding occurring in vowels. There is a known relationship between the implementation of lip rounding and both the vertical and the horizontal position of the tongue in vowels. Firstly, it is generally agreed that lip rounding in vowels is not realised uniformly across vowel heights (e.g., Catford, 1977; Lindau, 1978; Linker, 1982; Pasquereau, 2018). The higher the vowel, the smaller the degree of lip aperture. A high rounded vowel, such as [y], usually has a smaller lip opening than a lower rounded vowel, such as [ø]. This is probably due to mechanical reasons: it is hard to maintain close lip rounding when the jaw is opened. Secondly, accounts as early as Sweet (1877) indicate that lip rounding in vowels varies as a function of the frontness of the tongue. Two distinct configurations are generally described. One possibility is to form a small

lip aperture or a 'small tunnel' (Catford, 1988, p. 150) with the inner surfaces of the lips by bringing the lip corners in towards the centre horizontally. In this position, the lips have a 'pouted' configuration (Catford, 1988; Sweet, 1890). This type of rounding is associated with back vowels such as [u] and [o] and has been termed *inner rounding* by Sweet (1890), *horizontal* lip rounding by Heffner (1950) and *endolabial* by Catford (1988). Rounding of this sort involving a horizontal constriction of the space between the lips is predominantly associated with the contraction of the orbicularis oris muscle (Laver, 1980), as described in Section 3.1 (p. 76). The alternative lip rounding configuration is presented in relation to front rounded vowels like [y] and [ø], in which the lips are brought together vertically by closing the jaw (Ladefoged, 1971). While the side portions of the lips are in contact, a 'slit-like flat elliptical shape' (Catford, 1988, p. 150) gap is left in the centre. This configuration has been named *outer rounding* by Sweet (1890), *exolabial* by Catford (1988) and *vertical lip rounding* by Heffner (1950). The muscles implicated in this vertical compression of the lips include the inferior orbicularis oris and the mentalis, as well the raising of the jaw, as described in Section 3.1 (p. 76).

Ladefoged (1971) argued that a better pair of terms for the two types of rounding may be *lip rounding*, which would include lip protrusion, and *lip compression*. Lindau (1978) also distinguishes lip rounding, which for her is synonymous with lip protrusion, from lip compression. Accounts of lip protrusion vary, which could be due to the frequency of the different types of rounding across the world's languages. Heffner (1950) noted that 'protrusion of the lips is often a concomitant of horizontal lip rounding. It is much less frequently found with vertical lip rounding' (p. 98). Similarly, Laver (1980) remarked that lip protrusion is almost always accompanied by a certain degree of horizontal constriction of the space between the lips. However, he stressed that while substantial lip protrusion without horizontal constriction is physiologically possible, it is rare in the world's languages. He even went as far as to suggest that the articulatory parameter that all rounded vowels and labialised consonants have in common is the horizontal constriction of the inter-labial space. As a result, any labial articulation lacking a horizontal contraction would not be considered labialised by his view. Laver (1980) provided eight possible labial settings which deviate from a neutral lip position, which are

defined as combinations of horizontal and vertical expansion or constriction. He presented schematised representations of these settings, which have been recreated in Figure 3.2. As all eight settings may be accompanied by lip protrusion, there are in fact 16 possible deviations from the neutral labial setting. Laver (1980) explained that the most common labial setting in the world's languages is one involving horizontal constriction and vertical expansion with protrusion, which is the 'lip-rounded type of setting' (p. 38). The high frequency of this lip configuration may be due to the fact that back vowels are 'naturally' rounded and front vowels 'naturally' unrounded (Lindau, 1978). Furthermore, front rounded vowels are quite rare. As Mayr (2010) highlighted, out of the 562 languages studied in the *World Atlas of Language Structures*, only 6.6% are reported to have front rounded vowels (Maddieson, 2008). It is therefore not surprising that the most common lip action is the one associated with back vowels.

By combining the various phonetic accounts of the two main labialisation gestures described above, we define the main labialisation strategies as follows:

> Horizontal labialisation: associated with back vowels, the lips are pouted by drawing the lip corners together to form a small, rounded opening.
>
> Vertical labialisation: associated with front vowels, the lips come together by raising the bottom lip and closing the jaw, resulting in a small, slit-like opening.
>
> Lip protrusion: the lips are pushed forward to extend the vocal tract.

As Laver (1980) suggested, both horizontal labialisation and vertical labialisation may be accompanied by lip protrusion. We choose to avoid the somewhat loaded term *rounded* and employ instead the more phonetically neutral term *labialisation*, which can be applied to both consonants and vowels. We thus define labialisation as a secondary labial articulation in consonants and vowels resulting in a reduction of the overall lip area. By including lip area in the definition, we ensure that lip spreading, which would increase lip area, cannot be considered a possible labialisation strategy. However, we note that the size of the lip area of the two main labialisation gestures may vary. Presumably, horizontal labialisation with its small, rounded opening has a smaller lip opening to that of vertical labialisation with its slit-like

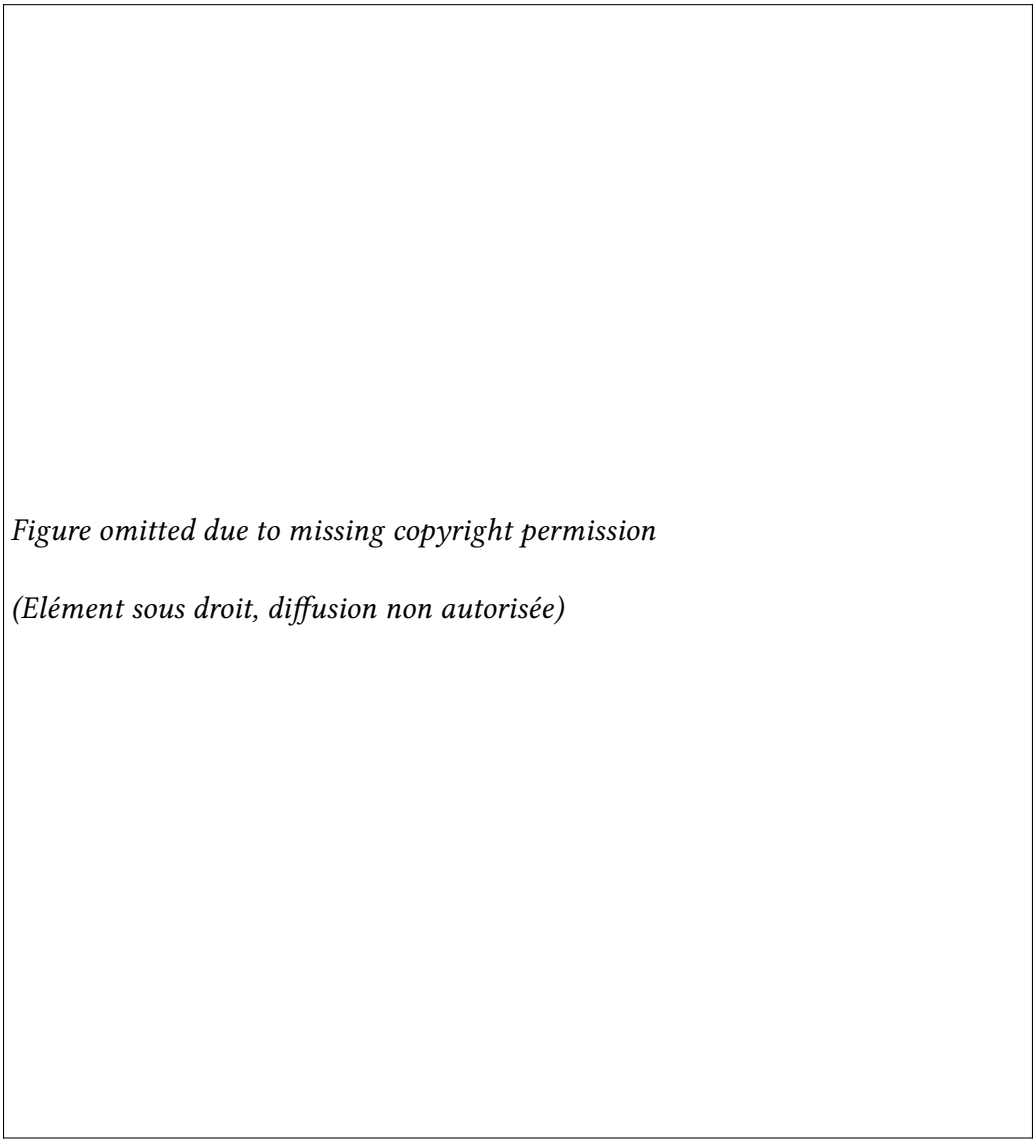



**Figure 3.2:** *Schematisation of possible lip settings according to Laver (1980, p. 37). All lip settings may be accompanied by lip protrusion. The outline of the neutral lip setting is indicated by a dashed line.* H – *Horizontal;* V – *Vertical;* E – *Expansion;* C – *Constriction.*

opening. Differences in the area of the lip opening are also suggested in the eight possible lip settings according to Laver (1980), schematised in Figure 3.2. Lip area differences may have acoustic consequences. Another component of labialisation, lip protrusion, which extends the size of the vocal tract, would also modify the acoustic output of speech. The acoustic effect of labialisation will be discussed in Section 3.5 (p. 84).

Detailed phonetic accounts of labialisation in consonants are surprisingly hard to come by. Most descriptions are phonological in nature in that they state whether or not a consonant is labialised, without going into details concerning the exact configuration of the lips. With regards to English consonants, it is generally agreed that the velar approximant /w/ and the post-alveolar sibilant fricatives /ʃ, ʒ/ are produced with labialisation. Descriptions of the post-alveolar approximant /ɹ/ may also include labialisation, which we described in more detail in Chapter 2 (p. 35). As /w/ is the semi-vocalic counterpart of /u/, it is assumed that /w/ and /u/ share the same labial properties. Some researchers have suggested that the post-alveolar sibilant fricatives have a different lip posture to that of /w/. Toda, Maeda, Carlen, and Meftahi (2003) studied lip patterns in both English and French using 3D facial motion-capture data from three subjects (one American, two French) producing nonsense words. They measured the front-back position of the lips and the approximate lip area in order to evaluate lip protrusion and 'rounding'. They suggested that 'labialisation' can be specified by two lip components. Both the post-alveolar fricatives /ʃ, ʒ/ and the velar approximant /w/ are produced with lip protrusion by all subjects, but the two consonant groups are opposed concerning lip area, /w/ being closed by rounding, unlike the post-alveolars which are described as 'open'. We can thus make a connection between Toda et al. (2003)'s two lip components in labialised consonants and the two possible lip shapes in rounded vowels previously described. Toda et al. (2003)'s description of /w/ is suggestive of horizontal labialisation, while /ʃ, ʒ/ without 'rounding' may be vertical labialisation. Brown (1981) also noted the similarity between labialisation in consonants and rounding in vowels. Inspired by Sweet (1877)'s description of 'inner' and 'outer' rounding, Brown also suggested that there are two lip gestures used for English labialised consonants: while /w/ is 'rounded', /ʃ, tʃ, ʒ, dʒ/ and /ɹ/ are 'protruded'.

## 3.4 LANGUAGE-SPECIFIC LABIALISATION

In most languages, vowels and consonants can simply be classified as labialised or non-labialised without requiring more detailed phonetic accounts because the different gestures involved in labialisation are not contrastive. However, as Ladefoged and Maddieson (1996) observe, it has been suggested that in some languages, there may be more than one distinct type of rounding gesture. The most well-known example is that of Swedish high front vowels, where /y/ and /ʉ/ have been shown to present contrastive lip configurations (e.g., Linker, 1982). Their labial postures are described by Ladefoged and Maddieson (1996) as horizontal rounding with protrusion, and compression, respectively. Similarly, although Japanese /u/ is generally considered unrounded, Nogita, Yamane, and Bird (2013) and Nogita and Yamane (2018) have suggested that /u/ actually involves protrusion without compression, and as a result, they argued that /u/ should be described as rounded. These language specific patterns suggest that in some cases, vowels that have traditionally been described as rounded, may not actually be produced with a rounded lip shape. Conversely, the somewhat restrictive terminology of *rounding* may have led to the labial gesture being overlooked in certain cases, such as the case of Japanese /u/. Labialisation, which is generally restricted to consonants, may thus be a more appropriate term than *rounding* for vowels as well as for consonants.

## 3.5 ACOUSTIC CORRELATES OF LABIALISATION

One of the very first things that phonetics students learn about the acoustics of vowels is that lip rounding lowers the frequency of formants. Basing his argument on Perturbation Theory (as discussed in Chapter 2, Section 2.8, p. 64), Stevens (1998) explained that lip rounding can be modelled as a decrease in the cross-sectional area at the open end of a uniform, unconstricted tube where one end is closed and the other end is open. However, he stressed that the downward shift in formants would apply not just to a uniform tube but to any arbitrary configuration that is open at one end, since there is a minimum in sound pressure and a maximum in volume velocity. Therefore, all articulatory-acoustic models should converge on the notion that any

articulation involving a decrease in lip area should result in the lowering of formants and indeed, to the best of our knowledge, they do.

However, vocal tract modelling has shown that the effect of lip rounding on formant frequencies is inextricably linked to the configuration of other articulatory parameters, namely the place and degree of the lingual constriction. In his influential *Acoustic Theory of Speech Production*, Fant (1960) showed that vowel formants can be accurately predicted by reducing the complexities of the vocal tract to a three-parameter, four-tube model, which was arguably one of the most important scientific breakthroughs in phonetics in the last century (Harrington, 2010). Fant (1960) presents nomograms to display the acoustic consequences of modifying the size and position of the lingual constriction as well as the degree of lip opening. An example of Fant's nomograms is presented in Figure 3.3 adapted from Fant (1989, p. 80). Figure 3.3 relates formant patterns to lingual constriction location with curves for five different degrees of rounding (from non-rounded curve 1 to very rounded curve 5). For these nomograms, the lingual constriction area is kept constant at 0.65 $cm^2$, which roughly corresponds to the narrow area of constriction in close vowels (Harrington & Cassidy, 1999). We first notice that changes to lip opening area predominantly affect F2 and F3, given the observable differences in F2 and F3 across the five lip area curves. Indeed, we know that the downward shift in formant frequencies caused by lip rounding in vowels particularly impacts F2 and F3 because they are affiliated with the front cavity (Vaissière, 2007). However, the nomograms show that formant frequencies are clearly affected by the varying horizontal location of the tongue constriction. When the tongue constriction occurs in the pre-palatal region (around 14 cm away from the glottis), lip area particularly affects F3. Conversely, F2 is predominately affected by lip area when the lingual constriction is more posterior (8-12 cm from the glottis).

Vocal tract modelling may give some indication as to why front and back rounded vowels are not produced with the same degrees of rounding, which we take to be synonymous with a horizontal contraction of the interlabial space. According to Vaissière (2011), what unites all 'focal' vowels is the merging of two adjacent formants in their acoustic profile. Although not a rounded vowel, as a starting point, we note that for focal [i], F3 and F4 need to be in

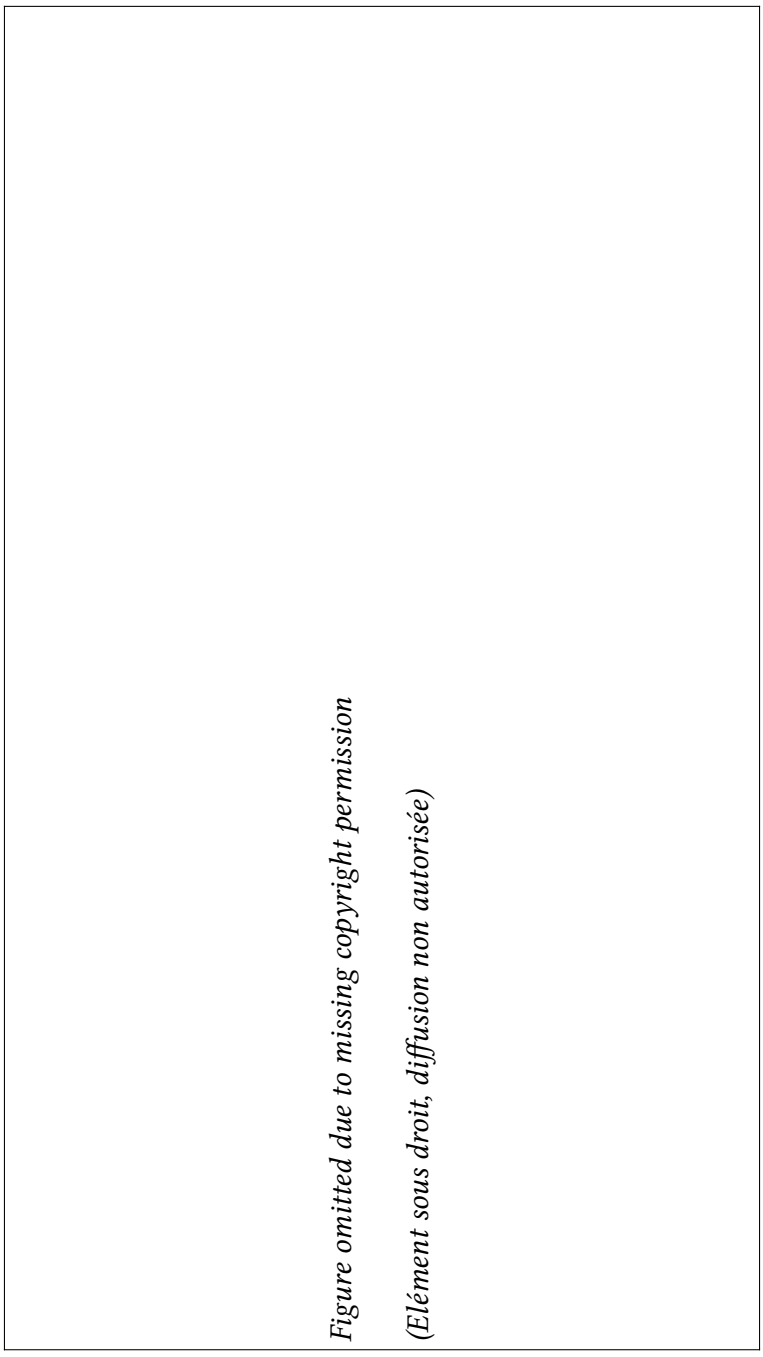



**Figure 3.3:** *Nomograms from Fant (1989, p. 80) for incremental values of lingual constriction location from the glottis to the lips with a constriction fixed at a narrow area of 0.65 cm$^2$. Curves 1-5 correspond to different lip areas from 8.00 cm$^2$ (no rounding) to 0.16 cm$^2$ (strong rounding). The points of formant merging are circled for* [i], [y] *and* [u].

close proximity (Vaissière, 2011). In Figure 3.3, the minimal distance between F3 and F4 for [i] (circled) occurs when the lingual constriction is positioned about 14 cm away from the glottis, i.e., in the pre-palatal region, and with the largest possible lip area (curve 1). For focal [u], F1 and F2 converge, which according to Vaissière (2007), is achieved with two strong constrictions: one at the palate and one at the lips. We find the smallest distance between F1 and F2 to occur with the smallest lip area curve presented in Figure 3.3 (curve 5), suggesting that close lip rounding is necessary in order to keep the distance between F1 and F2 maximally low when the lingual constriction is back (around 8 cm from the glottis). This suggestion was corroborated by Stevens (1998) who also noted that in the case of a backed tongue position, the condition of minimum F2 is achieved only if the lips are rounded and a narrow opening is formed. For a focal [y], formant merging occurs between F2 and F3 (Vaissière, 2011). The nomograms presented in Figure 3.3 indicate that when the lingual constriction occurs at least 10 cm from the glottis, some degree of clustering of F2 and F3 occurs in conjunction with all five lip area curves. However, the minimal distance between F2 and F3 does not occur in conjunction with the most lip rounding, but with a lip area of 2.0 $cm^2$ (curve 3). In the event of stronger lip rounding (i.e., with a smaller lip area), the distance between F2 and F3 increases. The frequency of F2 at the minimal distance between F3 and F2 for [y] is around 2 000 Hz, which is roughly the same frequency of F2 in focal [i]. The F1 of focal [i] and [y] do not greatly differ either. Therefore, what distinguishes focal [i] from focal [y] is the frequency of F3 (Vaissière, 2011). The proposal that the lips are not closely rounded for [y] in order to maintain the proximity of F3 to F2 was also suggested by Wood (1986). Similarly, Catford (1977) also explained that front vowels are usually 'exolabial', in order to avoid over-lowering the second formant and hence preserve their front quality. We conclude that while close lip rounding is needed for focal [u] to keep F2 maximally low, focal [y] maintains a minimal distance between F2 and F3 by avoiding close lip rounding. Wood (1986) argued that as the reported differences in lip articulation for [y] and [u] have always shown less lip rounding for [y], this difference can be considered a linguistic universal. The use of distinct labial configurations for front and back vowels may thus have acoustic and perhaps perceptual consequences, particularly in languages where rounding in vowels is contrastive, such as Swedish.

Although we have placed the point of formant merging for [u] to coincide with a back lingual constriction (around 8 cm from the glottis) where F2 is at its lowest in the nomograms in Figure 3.3, in reality, the lingual constriction is probably produced closer to the front of the vocal tract. The nomograms show that regardless of lip area, F2 descends quite rapidly as the lingual constriction moves from the pre-palatal region to the mid-palatal one (between 14-10 cm from the glottis) but then plateaus, suggesting that F2 is relatively insensitive to the location of the lingual constriction in the mid to back palatal region when the constriction is narrow. This is an example of a 'stable region' as proposed by Stevens (1989)'s Quantal Theory (as discussed in Section 1.5.1, p. 25), suggesting that quite large variations in the horizontal position of the lingual constriction would result in comparatively stable F1 and F2 frequencies (Harrington & Cassidy, 1999). French vowels are generally considered to be the closest you can get to cardinal vowel productions (e.g., Jones, 1972). For the French production of [u], the constriction is located around 11 cm from the glottis and the corresponding formant pattern is F1 = 250 Hz, F2 = 850 HZ, and F3 = 2 700 Hz (Savariaux, Perrier, & Orliaguet, 1995). A lingual constriction at this location actually occurs during F2's descent as the constriction moves back, prior to its low plateau. As a result, real F1 and F2 values are not as close together as indicated in the position of F1-F2 merging for [u] circled in Figure 3.3. It may be that for French /u/, rather than the merging of F1 and F2, the lowest possible concentration of energy is required, which is what we find with a lingual constriction 11 cm away from the glottis. Indeed, Ménard, Schwartz, Boë, and Aubin (2007) found that for French [u], F1 and F2 need to be minimally low and that focalisation of F1 and F2 occurs at the lowest position. We would like to stress that regardless of the place of the lingual constriction, the lowest possible F1 and F2 values always occur with the greatest degree of lip rounding (i.e., with the smallest lip area). The argument that [u] requires close lip rounding with a small lip opening is therefore still valid, whatever the front-back position of the tongue.

## 3.6 Motor equivalence and labialisation

The previous section indicated, perhaps somewhat implicitly, that multiple articulatory parameters may contribute to the frequency of a given formant. For example, we found that when the lingual constriction area is narrow, F2 lowering may be accomplished by positioning the lingual constriction further back (i.e., from a pre- to a mid-palatal position) or by increasing the degree of lip rounding. Increased lip protrusion would have a similar acoustic effect to the backing of the tongue because it would extend the cavity in front of the lingual constriction. It may therefore be possible to produce the same acoustic output using multiple articulatory configurations, which is sometimes known under the term *motor equivalence* (Perkell et al., 1993). As a result, trading relations may occur, whereby articulatory movements covary in order to keep a perceptually important acoustic cue constant (Brunner et al., 2011). The production of the vowel [u] is one such example. Perkell et al. (1993) found evidence to suggest that lip rounding and tongue-body raising are motor equivalent strategies for F2 lowering in American English /u/. They observed a trading relation between the degree of the labial and the lingual constriction: if one of these constrictions is too large (a property that tends to increase F2), the other constriction is adjusted accordingly (Perrier & Fuchs, 2015). Another possible trading relation may exist between lip rounding and the place of the lingual constriction for [u]. Savariaux et al. (1995) assessed how speakers of French behave when lip rounding for [u] is mechanically perturbed using a lip tube, which fixed the lip opening area at 4.9 $cm^2$. They found that 7 of their 11 speakers moved their tongue backwards to compensate for perturbation, which would have a lowering effect on F2.

Motor equivalence may result in the development of different production strategies for the same sound across speakers. One such case involving the lips may be the production of English /u/. A diachronic process of fronting of the /u/ vowel has been well-reported in Englishes worldwide, particularly in UK dialects (Fabricius, 2007; Ferragne & Pellegrino, 2010; Harrington, Kleber, & Reubold, 2008; Harrington et al., 2011). Acoustically speaking, /u/-fronting manifests itself as the raising of the second formant and is generally considered to be the result of the

lingual constriction being articulated at the front of the mouth. However, given the fact that F2 raising may also be the result of less lip rounding, Harrington et al. (2011) suggested that the F2 raising associated with /u/-fronting could be a result of either tongue-body fronting, lip unrounding, or a combination of both. In a study comparing the articulation and acoustics of /u/ in a variety of UK dialects, Lawson et al. (2019) found two distinct production strategies, which result in similar F2 values. In English and Irish speakers, the tongue was fronted and the lips were protruded. Conversely, Scottish tongue-body positions were located further back in the vocal tract but were accompanied by less lip protrusion. Although it would be tempting to consider these different production strategies a trading relation, as Lawson et al. (2019) pointed out, a difference in F1 between the two articulations does not make this possible. However, no correlation was observed between acoustic and articulatory frontness, suggesting that both the tongue and lips play a part in F2 lowering. This finding also highlights that tongue-body frontness should not be inferred from the acoustic signal alone.

## 3.7 CHAPTER CONCLUSION

We have shown in this chapter that the phonetic implementation of lip rounding is not as simple as what the binary phonological feature [± round] would suggest. Phonetic evidence has led us to call into question the appropriateness of the term *rounding* and its application to both consonants and vowels, as segments typically considered rounded, such as front rounded vowels, may not actually be produced with rounded lips. The somewhat restrictive label may have equally led to the labial gesture being overlooked in certain cases, such as Japanese /u/ (Nogita & Yamane, 2018; Nogita et al., 2013). We thus propose to use *labialisation* as a more phonetically neutral label, applicable to both consonants and vowels. We define labialisation as a secondary labial articulation, which results in a reduction of the overall lip area achieved via horizontal labialisation or vertical labialisation. Labialisation may also result in an increase in the length of the vocal tract when accompanied by lip protrusion.

Acoustic modelling has shown that modifications to the labial articulation have acoustic consequences. However, the lips combine with other articulatory configurations impacting

formant frequencies in different ways. For example, F2 is particularly affected by the size of the lip area when a narrow lingual constriction occurs in the mid to back palatal region. Conversely, lip area exerts greater changes to F3 when the narrow lingual constriction is positioned further front in the pre-palatal region. In order to produce a maximum acoustic and perceptual contrast between front and back labialised vowels, such as [y] and [u], their respective lip configurations may differ. Fant's nomograms would predict that for [u], the lowest possible F1 and F2 frequencies are attained with the smallest possible lip area, whereas for [y], as F2 needs to be as close as possible to F3, a larger lip area than the one for [u] is required. Horizontal labialisation, which is associated with back vowels, may therefore result in a smaller lip opening area than the vertical labialisation typical of front vowels. It seems then that the lips work in harmony with the tongue to form the necessary vocal tract configurations required for any given acoustic cue. Trading relations may occur across the different articulatory configurations which reciprocally contribute to a given acoustic cue. Reliance on one will result in less of another, and vice versa. Indeed, motor equivalence theory suggests that multiple vocal tract configurations may result in the same acoustic output. As a result, speakers may stray away from their habitual articulatory strategy for a given sound when an articulator is perturbed in some way in order to generate the expected acoustic/perceptual output. Motor equivalence phenomena such as these have provided possible answers to the long-debated question of whether speech production goals are articulatory or acoustic in nature. Motor equivalence argues in favour of acoustics. However, as Perrier and Fuchs (2015) pointed out, perturbation studies such as Savariaux et al. (1995) have also shown that if compensation is not possible, speakers will still prefer their usual vocal tract shape, suggesting that speech goals may have both articulatory and acoustic components.

# Summary and research questions

## 3.8 Summary and motivations

Although the articulation of English /r/ has been widely studied in rhotic Englishes, literature on non-rhotic Anglo-English is distinctly lacking, as we observed in Chapter 2. For one thing, there is a perception that Anglo-English /r/ is always produced with a tip-up tongue posture, although very little empirical evidence exists to back up this supposition. We first aim to fill this gap in the literature by accounting for the lingual gesture in Anglo-English /r/ in a larger cohort of speakers than in previous articulatory studies on Anglo-English /r/ (e.g., Delattre & Freeman, 1968) using Ultrasound Tongue Imaging (UTI). UTI has been used by the linguistics community for phonetics research since the 1960s (Gick, 2002a) but has gained in popularity as a technique in the last 20 years (Kochetov, 2020). When an ultrasound transducer is placed under the chin, ultra-high frequency sounds waves emitted from a crystal contained within the transducer travel through the tongue body tissue and are reflected back from the tongue surface in the form of echos (Stone, 2005). The echos are then converted into two dimensional images of the tongue surface, either sagittally or coronally. The fact that ultrasound cannot image bone or air means that it can only provide images of the tongue surface, and not, for

example, images of the jaw, pharyngeal wall or palate.[2] However, ultrasound is able to image the moving tongue in its near-entirety, producing high quality images, with good temporal resolution (30 fps or more) without causing discomfort or risk to the subject (Gick, Bernhardt, Bacsfalvi, & Wilson, 2008). As we are predominantly concerned with tongue shape, as opposed to its exact position in the vocal tract, UTI as a technique is well-suited to phonetic studies on English /r/ and has been used in a variety of previous studies including Heyne et al. (2018), Lawson et al. (2013) and Mielke et al. (2016) to name a few.

Although the labial articulation of /r/ has received less attention than its lingual one, we predict that by varying the degree of lip protrusion, speakers may attain similar acoustic outputs across the different tongue shapes for /r/. Systematic trade-offs have already been observed for English /r/ between the length of the front cavity and the length and size of the lingual constriction (Guenther et al., 1999) and a similar trading relation has been proposed between the sublingual space for tip-up /r/ and a more posterior palatal constriction for tip-down /r/ (Alwan et al., 1997). We predict that tip-down tongue shapes with negligible sublingual space will compensate with increased lip protrusion in order to maintain a large sized front cavity and therefore preserve a low third formant frequency in the resulting acoustic output.

To test to what extent lip protrusion contributes to the production of /r/, we will present lip, tongue and acoustic data from Anglo-English productions in both non-hyperarticulated and hyperarticulated speech. By eliciting hyperarticulated productions, it is hoped that speakers will be forced to enhance the discriminability of /r/, which will likely result in the lowering of the third formant, the most prominent acoustic cue for English /r/. If lip protrusion contributes to the lowering of F3, hyperarticulated /r/ may result in increased lip protrusion and therefore produce even lower F3 values than those observed in non-hyperarticulated productions of /r/. Lip protrusion will be measured using profile lip camera videos synchronised with both the ultrasound and the acoustic signal.

The labial articulation seems particularly pertinent to Anglo-English /r/ because labiodental variants are becoming increasingly common across England. Docherty and Foulkes (2001)

[2]The palate may be imaged indirectly by recording participants swallowing a bolus of water (Stone, 2005).

defined a change in progress whereby the labial component of Anglo-English /r/ is 'retained at the cost of the lingual articulation' (p. 183). They hypothesised that this change may be a 'function of the heavy visual prominence of the labial gesture' (p. 183). Underlying these claims are the following premises: firstly, that Anglo-English /r/ is produced with a labiodental articulation even when accompanied by a coronal gesture, and secondly, that this labiodental gesture is visually prominent. We intend to verify both of these claims with two further experiments. In Experiment 2, we will provide a detailed phonetic description of the labial gesture in Anglo-English /r/ by comparing it to that of /w/, whose articulation is unequivocally considered rounded. If /r/ is labiodental, its labial posture should differ considerably from that of /w/. The labial postures of /r/ and /w/ will be studied from front and profile lip camera data taken in Experiment 1. In a third experiment, we will assess the visual salience of the labial gesture in Anglo-English /r/ in a perception experiment. English participants will be presented with auditory-only, visual-only and congruous and incongruous audio-visual stimuli of /r/ and /w/. If the labial gesture of /r/ is labiodental and different to that of /w/, we expect the perception of /r/ to be enhanced with visual cues. Subjects may even be able to distinguish between /r/ and /w/ from the visual cues alone if the visual difference is particularly salient. In incongruous audio-visual stimuli in which auditory /w/ will be paired with visual /r/ and vice versa, visual capture may be anticipated if their respective visual cues are unambiguous and are more perceptually salient than the phonetic cues in the acoustic signal.

This thesis will not only contribute to the literature on the production of English /r/, but will have theoretical implications for the nature of speech perception in general, as well as for the role of visual speech cues in diachronic sound change.

## 3.9 Research questions

Given the observations gleaned from our review of the literature on the articulation of English /r/, on the phonetic implementation of labialisation and on multimodal speech perception more generally, the following research questions emerge:

1. Is the tip-up tongue shape typical of post-alveolar approximant /r/ in Anglo-English?
    (a) Is tongue shape subject to coarticulation with the following vowel as in other varieties of English?
2. How does lip protrusion contribute to the production of Anglo-English /r/?
    (a) Can lip protrusion enhance F3 lowering?
    (b) Is there a relationship between the degree of lip protrusion and lingual articulation?
3. Is Anglo-English /r/ produced with a labiodental articulation even in the cases where there is an observable tongue body gesture?
4. Is the labial posture for Anglo-English /r/ perceptually salient?

Part II presents two production experiments which will address questions 1-3. Part III concerns the perception of Anglo-English /r/ and will therefore address the final research question, question 4.

*Part* II

# Production of Anglo-English /r/*

*Portions of this work were published in King, H. & Ferragne, E. (2020). Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English. *Journal of Phonetics, 80*, 100978. doi:10.1016/j.wocn.2020.100978

# The articulation of Anglo-English /r/: Evidence from hyper- and non-hyperarticulated speech

# 4

## 4.1 Introduction

### *4.1.1 Aims and predictions*

Despite the abundance of articulatory studies on English /r/, Anglo-English remains largely unexplored, as our review of the literature in Chapter 2 indicated. There is an assumption that Anglo-English /r/ is produced with the tongue tip raised, which is perhaps due to the data presented in Delattre and Freeman (1968). However, with only three English subjects, their dataset can hardly be described as representative and Delattre and Freeman never claimed it to be so. We therefore aim to determine if the tip up tongue shape is indeed typical of /r/ in Anglo-English pre-vocalic /r/ by using a larger cohort of speakers. In non-rhotic Englishes, /r/ is produced in more retroflex-compatible contexts than in rhotic Englishes. Higher rates of retroflexion have been found in New Zealand English than American English. We intend to directly compare results from Anglo-English with the ones presented in Heyne et al. (2018) for New Zealand English and in Mielke et al. (2016) for American English. All

three studies utilise the same imaging technique (UTI) and speakers were recorded at a similar time (2016-2018). We will also assess whether similar phonetic factors to those observed in American English constrain tongue shape, focusing in particular on the impact of the following vowel. Retroflexion rates have been found to increase in the context of open-back vowels as opposed to close-front ones in American English, which is probably due to articulatory ease (Mielke et al., 2016). It has been shown in other varieties of English that the different tongue shapes associated with English /r/ do not differ with respect to the first three formants. We will assess whether the same can be said for Anglo-English. On a methodological level, there is currently no one technique that researchers use to classify tongue shapes for /r/ with UTI data, descriptions of which vary in detail. We aim to ensure our classification technique may be replicated by other researchers working with similar data. It is thus hypothesised that in Anglo-English:

**Hypothesis 1** /r/ is produced with higher rates of retroflexion than in American English.

**Hypothesis 2** /r/ tongue shapes are affected by coarticulation with the following vowel.

**Hypothesis 3** Different tongue shapes for /r/ result in similar formant values – at least up to F3.

After establishing how /r/ is articulated in Anglo-English with respect to its lingual component, we will turn our attention to the lips. As Chapter 2 indicated, it is clear that our understanding of the contribution of the lips to English /r/ acoustics is incomplete. While it is generally agreed that F3 is the main acoustic correlate for /r/, which is associated with front cavity resonances, we do not know to what extent the lips may influence /r/ acoustics. As Espy-Wilson et al. (2000)'s multi-tube models indicate, the addition of a separate lip protrusion channel would extend the front cavity and lower F3. However, do speakers actually put this articulatory strategy into practice? To test to what extent lip protrusion contributes to /r/, we will present data from both non-hyper- and hyperarticulated speech. If the final goal of speech movements is the correct perception of speech by the listener, the goal of hyperarticulation must be to enhance the discriminability of phonetic categories (as expressed by H&H Theory,

Lindblom, 1990). If the acoustic goal of English /r/ is indeed a low F3, hyperarticulated /r/ should reach even lower F3 values than those observed in non-hyperarticulated speech. If lip protrusion contributes to the lowering of F3, and therefore to the discernibility of /r/, we expect to find more lip protrusion in hyperarticulated speech than in non-hyperarticulated speech. We therefore postulate that:

**Hypothesis 4** Lip protrusion contributes to the lowering of the third formant of /r/.

The lips may also contribute to maintaining a stable acoustic output across different lingual articulations of /r/. As we pointed out in Chapter 2 (Section 2.8, p. 64), a trading relation between the tongue and lips may be a possibility. As the size of the sublingual space varies across tongue shapes, /r/ productions with little to no sublingual space may compensate by employing other articulatory manoeuvres which result in an increase in the size of the front cavity. Front cavity lengthening may be accomplished through a more posterior placement of the tongue, an extension of the sublingual space, or increased lip protrusion. Given the fact that labiodental articulations are rapidly gaining currency in England, we predict that Anglo-English /r/ has a labial component that may be related to the size of the sublingual space: articulations with little sublingual space, i.e., tip down bunched ones, may compensate with increased lip protrusion. Finally, if the trading relation between the sublingual space and lip protrusion exists, we may observe a larger degree of lip protrusion in bunched /r/ than in retroflex. In hyperarticulated speech, retroflexers may attain lower F3 values by increasing the size of the sublingual space (i.e., with more retroflexion), a strategy which would not necessarily be available to bunchers. We therefore predict that hyperarticulated bunched /r/ will be accompanied by more lip protrusion than hyperarticulated retroflex variants. If these arguments are valid, the following hypothesis can be derived:

**Hypothesis 5** A trading relation exists between the size of the sublingual space and the degree of lip protrusion, which manifests itself through a negative correlation between the two.

### 4.1.2 *Hyperarticulation*

In order to assess the contribution of the lips, articulatory and acoustic data from both non-hyperarticulated and hyperarticulated productions of /r/ will be presented. Speech communication is often characterised as a constant trade-off between ease of production and the successful transfer of information. For example, as described in Chapter 1 (Section 1.5.2, page 27), Lindblom's 'Hyper'- and 'Hypo'-articulation Theory (H&H Theory) states that speakers adapt their production according to the demands of the listener and the situation, which may account for the variable nature of the phonetics of speech (Lindblom, 1990). Thus, ease of articulation in the speaker is in direct opposition to the requirement for sufficient perceptual contrast for the listener (Bradlow, 2002). In fact, it has been shown that phonetic cues are often highly reduced in casual speech and may actually result in the loss of contrastive sound categories (Ernestus & Warner, 2011). Reduction may be related to the predictability of an utterance. Aylett and Turk (2004) found that phrase-medial syllables with high language redundancy (i.e., highly predictable from lexical, syntactic, semantic, and pragmatic factors) are shorter in duration than less predictable elements. They argued that the need for efficient information transfer while effectively expending articulatory effort leads to an 'inverse relationship between language redundancy and duration' (p. 31). This 'inverse relationship' improves communication robustness by spreading information more evenly across the speech signal, yielding a 'smoother signal redundancy profile' (p. 31). If the communicative situation places extra demands on the listener, we can expect the speaker to spontaneously adjust their articulatory patterns in order to produce speech that is 'clearer' (Bradlow, 2002). Types of speech that are produced with the goal of improving intelligibility are commonly referred to as *clear speech* or *hyper-speech* (Cooke, King, Garnier, & Aubanel, 2014). Speakers may adjust speech to accommodate to environmental demands when audibility is affected or perceived to be affected by the speaker. For example, speech is often modified in noisy conditions, known as *Lombard Speech* (Lombard, 1911) (e.g., Castellanos et al., 1996; Garnier, Heinrich, & Dubois, 2010; Junqua, 1993; Van Summers et al., 1988), or when addressed to a distant person (e.g., Cheyne, Kalgaonkar, Clements, & Zurek, 2009; Pelegrín-García, Smits, Brunskog, & Jeong, 2011). Speech modifications may also be

motivated by demands made by the target audience when they are perceived by the speaker to have intrinsically reduced comprehension, regardless of context (Cooke et al., 2014). Such instances include, but are not limited to, infant directed speech (e.g., Burnham, Kitamura, & Vollmer-Conna, 2002; P. K. Kuhl et al., 1997; Lindblom, Brownlee, Davis, & Moon, 1992; Stern, Spieker, Barnett, & MacKain, 1983), hearing-impaired directed speech (e.g., Bradlow, 2002; Howell & Bonnett, 1997; Picheny, Durlach, & Braida, 1985), speech addressed to non-native listeners (e.g., Scarborough et al., 2007; C. L. Smith, 2007; Uther, Knoll, & Burnham, 2007), machine directed speech (e.g., Burnham, Joeffry, & Rice, 2010a, 2010b; Oviatt, Levow, MacEachern, & Kuhn, 1996), and speech used when correcting (e.g., Beckford Wassink, Wright, & Franklin, 2007; Burnham et al., 2010a, 2010b; Schertz, 2013; Stent, Huffman, & Brennan, 2008).

Speech changes induced by environmental factors are primarily characterised by modifications to prosodic cues including increases in intensity, fundamental frequency and word duration (e.g., Castellanos et al., 1996; Garnier, Bailly, Dohen, Welby, & Loevenbruck, 2006; Van Summers et al., 1988). Some languages have even developed a whistled form of language in response to the necessity to communicate across very large physical distances (Meyer, 2005). In contrast, as Cooke et al. (2014) noted, listener-based speech modifications typically result in changes which may be considered as communicative strategies that help the listener to retrieve and decode phonetic cues. One such technique is exaggerated articulation, or hyperarticulation. On a segmental level, speakers have been shown to enhance phonetic contrasts between vowels and between consonants. Enhancement strategies may include increases to the vowel space, exaggerated jaw and lip movement, and changes to length contrasts in vowels and voicing contrasts in consonants (a review of known speech modifications is presented in Cooke et al., 2014).

Speech has been found to be hyperarticulated in computer- compared with human-directed speech (Burnham et al., 2010a), particularly in speech following recognition errors (Maniwa, Jongman, & Wade, 2009; Oviatt et al., 1996; Schertz, 2013). If only one segment is incorrectly identified, or is likely to be misunderstood, speakers may limit and target their adaptations to that particular segment in subsequent productions (Schertz, 2013), i.e., *targeted hyperar-*

*ticulation*.[1] A number of studies have elicited targeted hyperarticulation by employing an experimental paradigm in which participants interact with a simulated automatic speech recogniser and receive text feedback about what the programme 'recognised'. Stent et al. (2008) found that American English speakers make repairs after recognition errors and that hyperarticulation increases after evidence of misrecognition and then gradually decays in the absence of further misrecognitions: speakers' pre-error speaking style usually returns 4–7 utterances after evidence of misrecognition. The authors found repairs to typically include the use of canonical forms rather than reduced or assimilated ones, e.g., the flapping of /t/ was modified to [t]. In Schertz (2013), participants interacted with a simulated automatic speech recognition system and had to repeat words which were incorrectly identified. Target words included voiced and voiceless plosive onsets (e.g., *pit, bit*). More extreme voice onset time (VOT) values were elicited by an incorrect computer recognition in which the error was a minimal pair in voicing with the target plosive (e.g., subject reads *bit*, computer responds with 'pit'). However, when the computer gave an open-ended request for repetition (e.g., subject reads *bit*, computer responds with 'What did you say?'), hyperarticulation did not occur. In Buz, Tanenhaus, and Jaeger (2016), subjects were recorded interacting with a simulated human partner over the web. Subjects were asked to say one of three words which appeared on a screen and were informed that their partner would select the word they understood from the three options. Target words contained voiceless plosive onsets. The results indicate that speakers hyperarticulate the target word when a voiced competitor is present and that the size of the hyperarticulation effect was nearly doubled when simulated partners occasionally misunderstood the word.

The results from previous studies suggest that speakers make judgements based on the 'perceived communicative success' (Buz et al., 2016) of their utterances and adapt their speech accordingly. The properties of speech that speakers modify in order to improve the intelligibility of their speech do not all occur at the same time and under the same conditions (Stent et al., 2008). As previously discussed, environmentally-driven modifications tend to occur globally in order to improve audibility. In contrast, listener-oriented adaptations tend to occur more locally with the goal of enhancing segmental distinctiveness. As a result, hyperarticulation may be

[1] Other labels have also been employed including *contrastive*, *focal* and *localised* hyperarticulation.

considered to be a gradient process. Possible enhancement strategies may arise from speakers learning from their experience of the most effective techniques to convey their intended message in a given situation. Indeed, some studies have shown that spontaneous speech adaptations improve intelligibility in listeners (e.g., Junqua, 1993; Krause & Braida, 2003), although not all reported enhancement strategies have necessarily proven to be beneficial (see Cooke et al., 2014, for a review of the perceptual effects of speech adaptation).

While previous studies have been interested in how and why speech enhancement modifications occur, we intend to elicit adaptive behaviour in order to answer a specific research question relating to the phonetic implementation of a particular segment, English /r/. If the final goal of speech movements is the correct perception of speech by the listener and if the acoustic goal of /r/ is a low third formant, articulatory enhancement should result in further F3 lowering. We will assess which articulatory parameters are available to speakers in order to enhance English /r/ by eliciting targeted hyperarticulation at a segmental level. Our methodology will draw on the results from previous studies, which have indicated that the highest rates of targeted hyperarticulation occur in computer- rather than human-directed speech, in speech repairs directly following recognition errors and in the 4-7 utterances following the initial error.

## 4.2 Methodology

### *4.2.1 Procedure*

In order to elicit targeted hyperarticulation specifically at a segmental level, we engaged speakers in error resolution with a simulated speech recognition programme. Speakers were deceptively informed that the aim of the experiment was to test a new automatic 'silent speech' reader, which used information from speech movements to recognise the words they say without referring to the auditory signal. They were told that the silent speech reader was having difficulties with certain speech sounds, and that the aim of the recordings was to test the programme on these sounds. However, the sounds of interest were never explicitly revealed to

subjects. The experiment was divided into two parts. During the first, speakers were informed that the computer had access to both visual and auditory cues from their speech. During this part, the programme correctly 'identified' every word uttered, which provided us with baseline, non-hyperarticulated productions of /r/. During the second part, participants were informed that the audio would be 'turned off' and that the programme would only have access to visual speech information from their lingual and labial movements. During this second part, the computer 'incorrectly' identified one third of the stimuli. Whenever computer errors occurred, participants were instructed to repeat the word to try to 'make the computer understand'. Each 'incorrectly' identified word was repeated two more times in a row, the first of which elicited the same 'incorrect' response before being 'correctly' recognised. Recording sessions lasted no longer than 30 minutes and the stimuli were presented in a randomised order. By telling participants that the programme could not hear them, it was hoped that articulatory adaptations would be made locally at a segmental level, rather than across the entire word, which may have involved prosodic changes. Participants were told to use their normal speaking voice throughout the recording session.

The target word and computer feedback were presented to the participant, who was seated in a sound-attenuated room, on a computer screen. The participant first saw the target stimulus, e.g., *reed*, and the experimenter initiated the recording, which produced a beep sound in the sound-attenuated room, signalling to the participant to say the word on the screen. The participant then saw the message 'processing...please wait', which gave time for the experimenter, who was seated in an adjacent control room, to select the appropriate computer response. There were three possible computer feedback responses:

1. Recognition not possible: 'Word not recognised, please wait.'
2. Incorrect identification: 'Did you say weed?'
3. Correct identification: 'Did you say reed?'

Although the simulated feedback responses had been pre-determined, the first possibility (i.e., 'Word not recognised, please wait') was included in case a subject made a mistake, in which

case a target word could be repeated without jeopardising the believability of the simulated programme. We had originally considered using a technique in which the simulated feedback response was automatically presented to the participant as soon as the experimenter had pressed the stop button. However, pilot testing indicated that subjects very quickly realised that the automatic speech reader was simulated if they made a mistake or did not respond in time but the programme was still able to correctly 'recognise' the word they had been asked to say. Pilot testing also indicated that if the computer recognition feedback was simply presented to the participant directly after having produced the word, some participants paid little attention to the computer feedback response, focusing instead on the words they were asked to say. In order to elicit targeted hyperarticulation of /r/, it was vital that participants believed that their production of /r/ was the source of computer misrecognitions. As a result, after each recording, the participants were asked to confirm whether the computer had correctly identified the word they had just said, such as in the following: 'Did you say reed?'. Participants then responded with yes or no, which they were told would trigger the programme to either move on to the next word in the word list or repeat the original target word if automatic recognition was incorrect. A schema depicting the order of possible computer responses is presented in Figure 4.1.

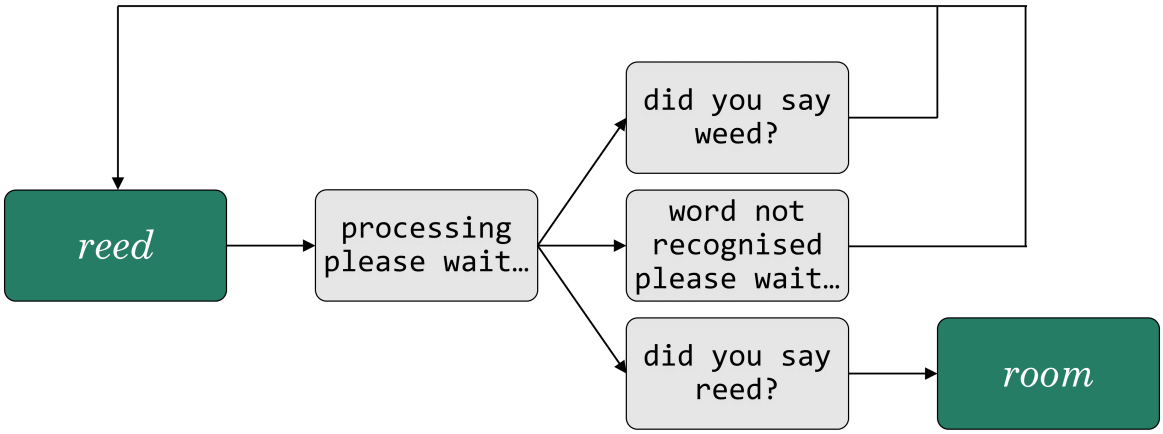



**Figure 4.1:** *Possible responses from the simulated automatic silent speech reader (presented in grey) after a target word (presented in green), here the target word is* reed*). If the computer feedback was a misrecognition (here,* weed*), the target word was repeated two more times in a row, the first of which elicited the same 'incorrect' response (here,* weed*). The second repetition resulted in a correct recognition, after which a new target word was presented (here,* room*).*

In order to ensure the believability of the simulated programme, a simulated programme interface (presented in Figure 4.2) was created and presented to speakers on a separate screen throughout the recordings. Fake on/off buttons were shown next to the words 'audio', 'video' and 'ultrasound'. Just before the second 'silent speech' part started, the experimenter 'turned off' the audio by clicking on the corresponding fake button.

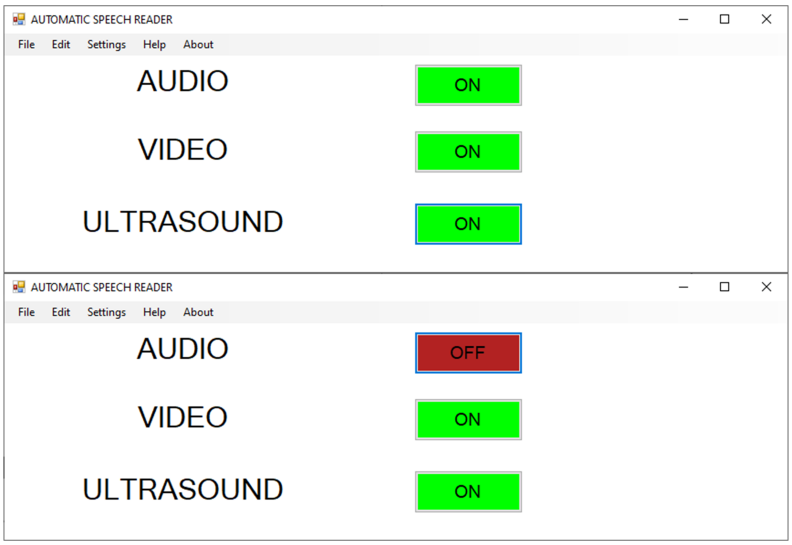



**Figure 4.2:** *Simulated 'Silent Speech Reader' interface presented to subjects during the non-hyperarticulation (top) and hyperarticulation (bottom) session.*

### *4.2.2 Stimuli*

Stimuli comprised of nine /r/-initial monosyllabic words followed by the vowels FLEECE, GOOSE, KIT, DRESS, TRAP, STRUT[2], THOUGHT, LOT. Fillers were /w/-initial words followed by the same monophthongs. In the non-hyperarticulated session, all target words were 'correctly' identified by the simulated programme. To ensure believability, one repetition per item was recorded in the first session. For the second hyperarticulated session, /r/ productions in the words *reed*,

[2]Some speakers, particularly those from the north and the midlands of England, may not present the FOOT-STRUT split. As a result, we expect the STRUT vowel to be variable with linguistic Northerners likely producing the near-close near-back round FOOT vowel rather than the open-mid back unrounded STRUT vowel.

*red*, and *room* were 'incorrectly' identified as 'w' and 'l' (e.g., *red* was identified as 'wed' or 'led'). When an 'incorrect' response was given, the original word was repeated two more times. The same method was used for /w/-initial filler words, where /w/ productions were mistaken for 'r' or 'l'. A total of 24 productions of /ɹ/ were recorded in the second, hyperarticulated session. Stimuli were presented to subjects in a semi-randomised order. In the hyperarticulated session, misrecognitions were never followed by more than four correct recognitions of /ɹ/ or /w/ to ensure targeted hyperarticulation was maintained throughout the session (based on results from Stent et al., 2008, as discussed in Section 4.1.2). A complete list of stimuli is presented in Table 4.1.

| Target word | Lexical set | Transcription | Misrecognition I | Misrecognition II |
|---|---|---|---|---|
| *reed* | FLEECE | /ɹiːd/ | 'weed' | 'lead' |
| *red* | DRESS | /ɹɛd/ | 'wed' | 'led' |
| *room* | GOOSE | /ɹuːm/ | 'womb' | 'loom' |
| *reap* | FLEECE | /ɹiːp/ | | |
| *ring* | KIT | /ɹɪŋ/ | | |
| *rack* | TRAP | /ɹæk/ | | |
| *run* | STRUT | /ɹʌn/ | | |
| *raw* | THOUGHT | /ɹɔː/ | | |
| *rot* | LOT | /ɹɒt/ | | |
| *weed* | FLEECE | /wiːd/ | 'reed' | 'lead' |
| *wed* | DRESS | /wɛd/ | 'red' | 'led' |
| *womb* | GOOSE | /wuːm/ | 'room' | 'loom' |
| *weep* | FLEECE | /wiːp/ | | |
| *wing* | KIT | /wɪŋ/ | | |
| *whack* | TRAP | /wæk/ | | |
| *one* | STRUT | /wʌn/ | | |
| *war* | THOUGHT | /wɔː/ | | |
| *what* | LOT | /wɒt/ | | |

**Table 4.1:** *List of stimuli and fillers. To elicit targeted hyperarticulation, one third of words were 'incorrectly' recognised by a simulated automatic speech recognition programme. Simulated computer misrecogntions are presented. For non-hyperarticulated productions, all words were correctly identified. Phonological transcriptions are based on Standard Southern British English.*

### 4.2.3 Equipment

Data were collected in a sound-attenuated room in the Clinical Audiology, Speech and Language Research Centre at Queen Margaret University, Edinburgh. Simultaneous audio, ultrasound and lip camera data were obtained using the 'Record Ultrasonic plus Video' option in Articulate Assistant Advanced (AAA) software (Articulate Instruments Ltd., 2014) (version 2.16.16). Although the three data signals are collected as independent streams, AAA permits all channels to be started and stopped with the click of a single button and these channels are synchronised within the software. Stimuli were presented to the participant using AAA. Each word appeared at the top of the recording screen, as shown in Figure 4.3. The entire screen was visible to the experimenter who was seated in the control room, while only the top portion containing the stimulus was visible to the speaker during recording sessions.

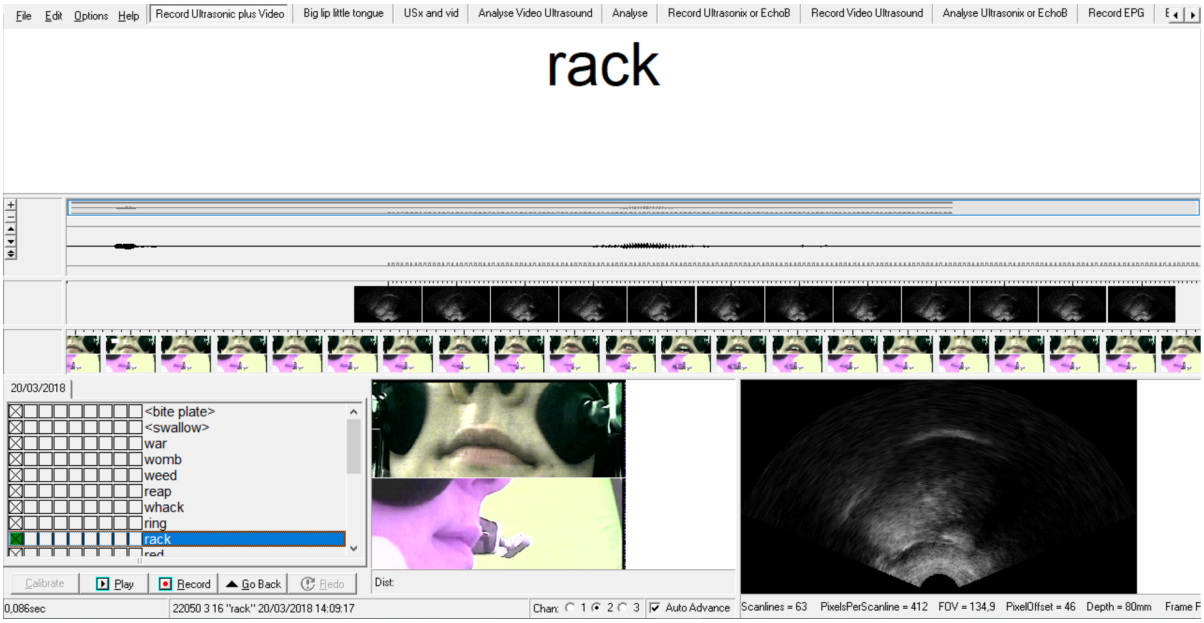



**Figure 4.3:** *Example screen display during recording sessions. Although the experimenter saw the entire screen in the control room, only the top portion of the screen containing the stimulus and the simulated computer feedback was visible to the speaker.*

Ultrasound tongue images were recorded at a rate of circa 121 frames per second (fps) using a high-speed SonixRP ultrasound system with a 5 MHz 10 mm radius microconvex probe. The probe was positioned underneath the jaw and angled so that the tongue tip was maximally visible. The probe was fixed in place relative to the speaker's head using an Ultrasound Probe Stabilisation Headset (Articulate Instruments Ltd., 2008), presented in Figure 4.4. Efforts were taken to ensure that speakers did not wear the headset for more than 30 minutes. Two NTSC micro-cameras were attached to the headset, capturing front and profile lip videos[3] at a rate of circa 60 fps. An Audio-Technica AT803 directional clip-on microphone was clipped to the side of the headset. Audio files were digitised as LPCM mono files with a 22 050 Hz sampling rate and 16-bit quantization. Technical details concerning this particular ultrasound system and associated video and audio synchronisation are described in Wrench and Scobbie (2016).

We recorded each speaker swallowing water in order to obtain an outline of the palate (Epstein & Stone, 2005). Speakers were also recorded biting on a plastic bite plate, which was used to image each speaker's occlusal plane (Lawson et al., 2019). The palate and occlusal plane were subsequently traced in AAA.

[3]Only data from the profile video camera will be presented in this first experiment.

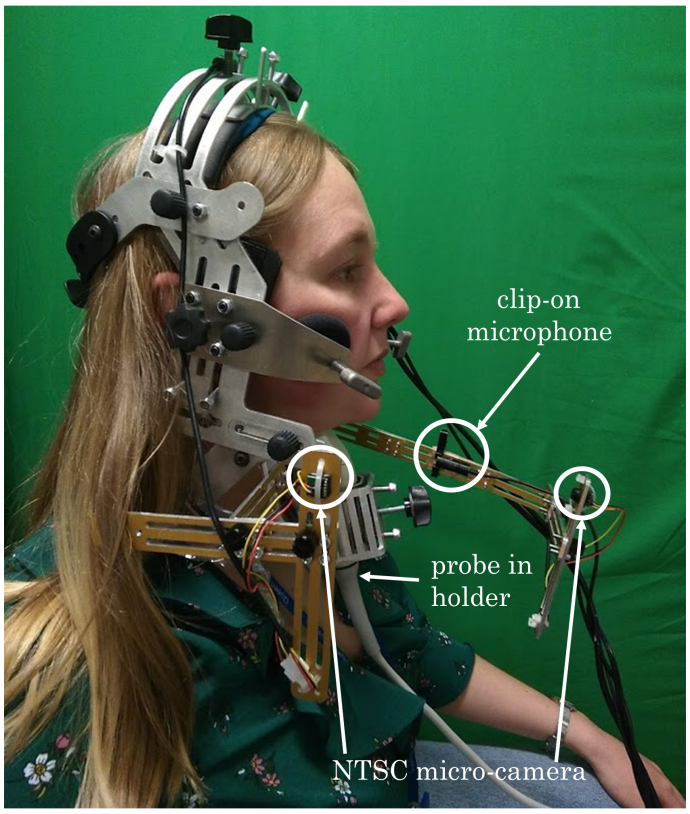



**Figure 4.4:** *The author demonstrating the use of the Ultrasound Stabilisation Headset with clip-on microphone, front and profile NTSC micro-cameras and ultrasound probe in holder.*

### 4.2.4 Participants

29 native speakers of Anglo-English were recorded at Queen Margaret University, Edinburgh. Speakers were recruited through advertising on the university's Research Recruitment Digest communications service. Participants self-identified as speaking with an English accent and we made sure that this was indeed the case by conversing with participants before recording them. Before participating, speakers signed an informed consent form (presented in Appendix A.1) and completed a background questionnaire (presented in Appendix A.2). Ethical approval had previously been obtained from Queen Margaret University Research Ethics Committee. Subjects were financially compensated £20 for their participation.

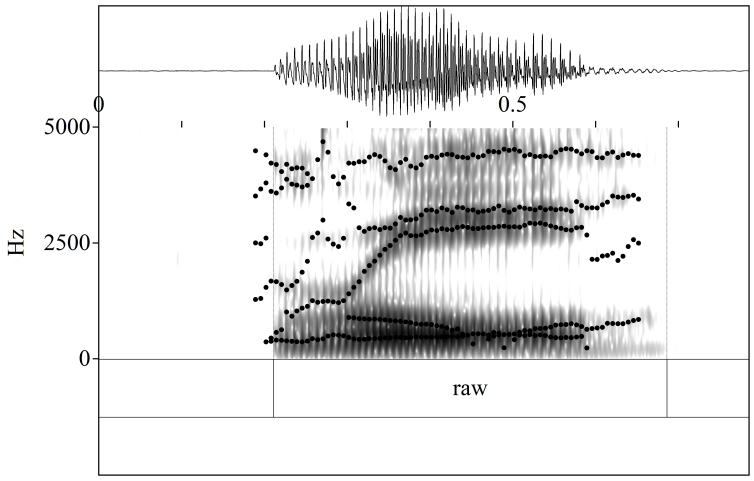
Some speakers' data were excluded due to ultrasound data visualisation issues ($n$=4) and one English-Punjabi bilingual was excluded because Punjabi also has retroflex consonants in its inventory. We present data from the remaining 24 speakers (22F, 2M) aged between 18 and 55 (mean = 30.08 ± 11.26) who come from all over England (south west: $n$=1; south east: $n$=6; midlands: $n$=3; north west: $n$=7, north east: $n$=7). 19 speakers had lived in Scotland for at least a year. We would have preferred a more balanced sample with regards to speaker sex. One reason for the disparity between the sexes in the present dataset is no doubt due to the fact that Queen Margaret University has a particularly high proportion of women in its student population, at 76%, as outlined in the university's most recent Gender Action Plan (Queen Margaret University, 2017). The inclusion of the word *war* in the stimuli allowed us to classify the participants as rhotic and non-rhotic. All speakers were non-rhotic apart from the one speaker from the south west of England, where rhotic accents do indeed occur (Wells, 1982), although they are becoming less rhotic (Trudgill, 1999a). Incidentally, this subject is one of the oldest speakers in the dataset (54 years old). Table 4.2 presents demographic information for all 24 speakers. Languages spoken at a high level, i.e., beyond intermediate (B2), have been indicated.

| Subject code | Sex | Age | Origin | Languages | >1 year in Scotland |
|---|---|---|---|---|---|
| 02 | F | 22 | north west | | ✓ |
| 03 | F | 22 | north east | | |
| 04 | M | 53 | north east | German (advanced) | ✓ |
| 05 | F | 22 | south east | | |
| 07 | F | 22 | north east | | ✓ |
| 08 | F | 26 | north west | | ✓ |
| 09 | F | 21 | south east | Cantonese (bilingual) | |
| 10 | M | 44 | north west | | ✓ |
| 11 | F | 29 | midlands | | ✓ |
| 12 | F | 20 | north west | | |
| 13 | F | 54 | south west | | ✓ |
| 14 | F | 23 | south east | | ✓ |
| 15 | F | 25 | north west | | ✓ |
| 16 | F | 25 | south east | | ✓ |
| 17 | F | 27 | north east | | ✓ |
| 18 | F | 23 | south east | | ✓ |
| 19 | F | 28 | north east | | ✓ |
| 21 | F | 41 | north west | | ✓ |
| 22 | F | 23 | north east | | ✓ |
| 23 | F | 33 | midlands | | ✓ |
| 25 | F | 55 | midlands | | ✓ |
| 27 | F | 37 | north east | | ✓ |
| 28 | F | 29 | north west | | ✓ |
| 29 | F | 18 | south east | | |

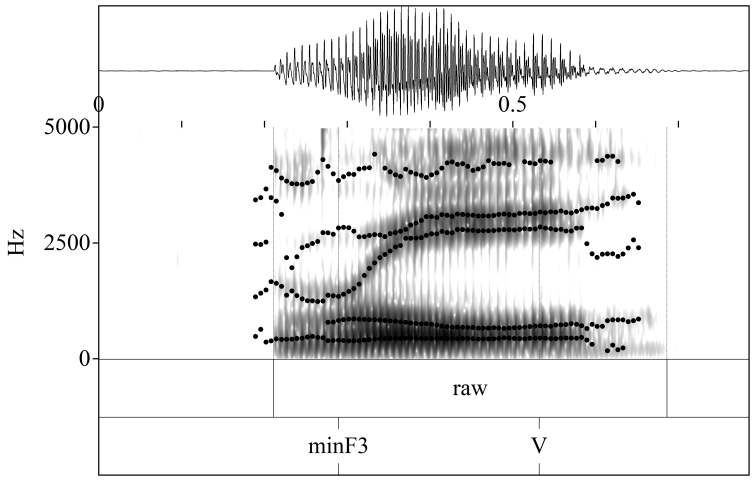
**Table 4.2:** *Participant demographics from production experiments. Languages spoken at an advanced level or above have been included.*

### 4.2.5 Acoustic analysis

The acoustic data were exported as wav files from AAA and analysed in Praat (Boersma & Weenink, 2019). Determining the point at which to segment /r/ from the following vowel is challenging. Although Lawson, Stuart-Smith, Scobbie, Yaeger-Dror, and Maclagan (2010) suggested that for post-vocalic /r/, the most reliable means to determine the dividing line between the two is by considering amplitude changes, in our prevocalic /r/ data, we observed large amounts of amplitudinal variation both within and across speakers. We were therefore unable to find a sufficient technique that could be applied to all speakers. As a result, /r/ and the following vowel were manually annotated as a whole. Praat's Burg algorithm was used to obtain formant values. For each recording, formant parameters were manually adjusted in order to reach an optimal match between formant estimation and the underlying spectrogram. This generally involved adjusting the ceiling of the formant search range in Hertz using the 'Maximum Formant' setting in Praat. For example, Figure 4.5a shows the waveform, spectrogram and the formant estimation set at 5 000 Hz for the word *raw* produced by a male speaker. The formant tracks evidently do not match the underlying spectrogram, particularly towards the middle of the vowel. As a result, the ceiling of the formant search range was adjusted to 4 500 Hz, yielding much more adequate formant tracks (as presented in Figure 4.5b). Once the parameters were optimised, the formant listing for the portion corresponding to /r/ and the following vowel was opened in Praat (under 'Formant' → 'Formant listing'). The minimum F3 value (as in Guenther et al., 1999) was found within this formant listing and the point corresponding to the minimum F3 value was labelled (as depicted in Figure 4.5b). A point during a steady state of the vowel following /r/ avoiding any obvious transitions to and from the surrounding consonants was selected and labelled. The first three formants (F1-F3) were then extracted at these two points, i.e., at minimal F3 for /r/ and during a steady state of the following vowel.




**(a)** *Prior to formant parameter optimisation. Ceiling of formant search range set to 5 000 Hz.*




**(b)** *After formant parameter optimisation. Ceiling of formant search range set to 4 500 Hz. The first three formants were extracted at the resulting point of minimal F3 (labelled 'minF3') and a steady state of vowel (labelled 'V').*

**Figure 4.5:** *Waveform, spectrogram and formant estimation for the word* raw *produced by a male speaker* **(a)** *before formant parameter optimisation and* **(b)** *after formant parameter optimisation.*

### 4.2.6 Ultrasound analysis

One ultrasound frame was selected per recording depicting the maximal constriction of the anterior lingual gesture for /r/ prior to any obvious movement into the following vowel. This was achieved by holistically examining the raw ultrasound images one by one in a sequence. For each selected image, a spline was fitted to the visible surface of the midsagittal tongue using the edge-detection algorithm in AAA. A preliminary step to the edge detection process requires selecting the upper and lower limits between which the algorithm may detect a bright surface (Lawson et al., 2013). The upper and lower limits are set to remove traces of the hard palate and bright areas resulting from muscle structures inside the tongue (Lawson et al., 2013), which can be observed in the left image of Figure 4.6. After tracing the upper and lower limits (indicated by the green lines in Figure 4.6), a spline was roughly traced by hand around the midsagittal tongue contour (presented in pink in the middle image in Figure 4.6). AAA's edge-detection algorithm was then implemented by pressing the 'Snap-to-fit' button. In the parts of the contour that have a good edge, the spline appears as a solid line. As the right image in Figure 4.6 indicates, automatic edge detection removes parts of the spline at the extreme right and left of the image where no clear tongue surface exists in the original ultrasound image. Occasionally, the algorithm may miss certain areas of the tongue contour, particularly around the tongue tip due to shadowing from the jaw. In these cases, splines were manually corrected. Corrections were often achieved through a holistic examination of the ultrasound frames occurring before and after the selected frame, which generally allowed for more accurate tracking of the tongue tip, rather than relying on the one static ultrasound image selected.

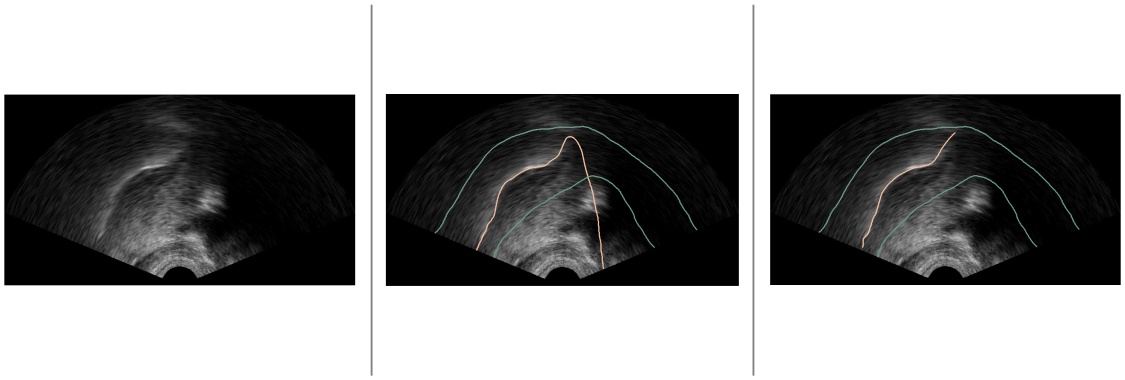


**Figure 4.6:** *Automatic detection of the midsagittal tongue contour in ultrasound data. (left) Raw ultrasound image depicting the visible surface of the midsagittal tongue represented by the lower edge of the bright white line; (middle) Upper and lower limits (green lines) and tongue surface spline (pink line) hand-traced for automatic edge detection; (right) after automatic edge-detection spline fitting.*

### Imaging the occlusal plane

A reference spline was fitted to each speaker's occlusal (i.e., bite) plane, which was imaged using a bite plate.[4] Imaging a speaker's occlusal plane improves interpretation of tongue position and inter-speaker comparison (Lawson et al., 2019). Bite plates are made from 2mm thick medical grade plastic and vacuum-moulded around a standard template (95x40 mm) (Lawson, Stuart-Smith, Scobbie, & Nakai, 2018), an image of which is presented in the top left image in Figure 4.7. When the bite plate is placed in the mouth (top right image of Figure 4.7), a vertical ridge located near the middle of the bite plate rests against the front of the upper incisors. Participants were recorded biting on the bite plate and were asked to press their tongue against the underside. The resulting flat surface of the tongue pressed against the bite plate allows for the identification of a flat plane in the ultrasound video image (Lawson, Stuart-Smith, Scobbie, & Nakai, 2018), which is visible in the bottom left ultrasound image presented in Figure 4.7. A reference spline was fitted to this plane (as presented in the bottom right image of Figure 4.7), which was used to rotate all subsequent splines to a quasi-horizontal position. Figure 4.8 depicts

[4]Bite plates were kindly provided by the Clinical Audiology, Speech and Language Research Centre at Queen Margaret University, Edinburgh.

our rotation technique: all contours are rotated so that the occlusal plane (green line) tracing is horizontal. An alternative technique is to adjust the probe-to-chin angle using the stabilising headset before recording so that the image of the occlusal plane is parallel to the upper and lower edges of the video pane, as described in Lawson et al. (2019). We decided to rotate the tongue surface splines at the post-processing stage because we found that in some speakers, adjusting the probe angle reduced the visibility of the tongue tip, which was of particular importance in the current study.

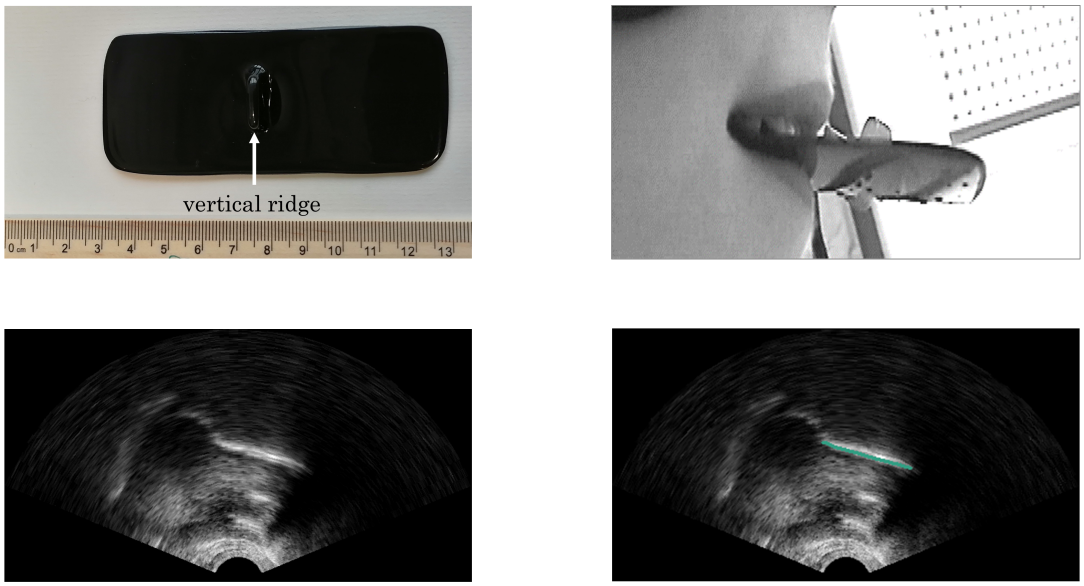



**Figure 4.7:** *Imaging and detecting the occlusal plane with a bite plate. (top left) Plastic bite plate; (top right) Subject biting on bite plate; (bottom left) Resulting ultrasound image depicting flat surface of tongue against the underside of the bite plate; (bottom right) Reference spline tracing used to rotate all subsequent splines to a quasi-horizontal position.*

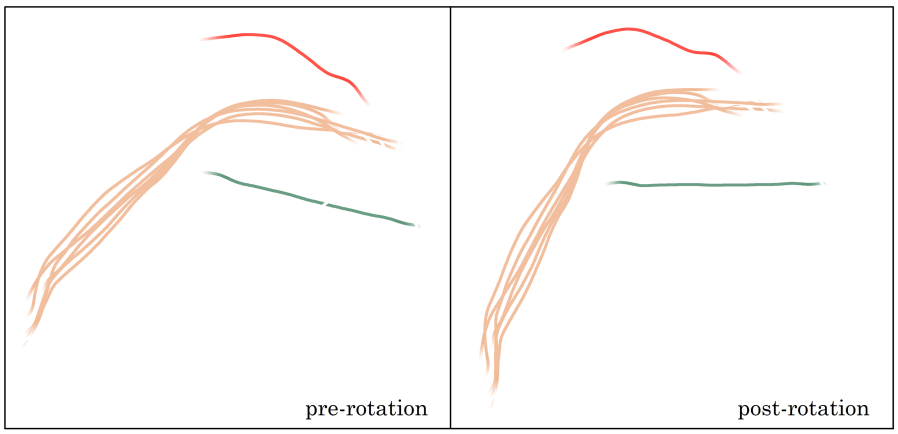



**Figure 4.8:** *Example of rotation of splines to the occlusal plane. The tongue tip is on the right. The hard palate is traced in the top curve. All contours are rotated so that the occlusal plane (bottom line) is horizontal.*

### Identifying tongue shapes

Both the raw ultrasound images and the tongue splines rotated to the occlusal plane were used to classify tongue configurations for /r/ on a continuum largely inspired by the one presented in Lawson et al. (2013) for Scottish English, which depicts four distinct shapes: Mid Bunched, Front Bunched, Front Up and Tip Up (pp. 199–200), as presented in Section 2.4 (p. 41). Our classification differs in that it includes a fifth configuration: an 'extreme' sublaminal retroflex involving curling up of the tongue tip, which has previously been associated with Anglo-English (as discussed in Section 2.4). The classification originally proposed by Lawson et al. (2013) grouped the curled up and the non-curled up tip up /r/ together. Ultrasound images give some indication of the curling up of the tongue tip, which we describe below. However, we do not know to what extent the identification of these articulations is constrained by speaker anatomy. In some cases, it is possible that the jaw shadow obscures the tongue tip, which would make visualising sublaminal retroflexion challenging. It is therefore possible that the number of curled up articulations is underestimated in our analysis. The articulations of each

configuration in our classification are described below,[5] and Figure 4.9 presents raw ultrasound images of typical examples of each configuration from our dataset.

**Mid Bunched (MB)**: the middle of the tongue is raised towards the hard palate, while the front, blade and tip are low.

**Front Bunched (FB)**: the front of the tongue has a distinctly bunched configuration which results in a dip in the tongue's surface behind the bunched section. The tip and blade remain lower than the rest of the tongue front.

**Front Up (FU)**: the front, blade and tip are raised and the tongue surface forms a smooth convex curve.

**Tip Up (TU)**: the tongue tip is pointing up resulting in a straight and steep tongue surface.

**Curled Up (CU)**: the overall tongue shape is concave and the tip is curled up. Curling up of the tongue tip results in a near-parallel orientation of the tongue surface to the ultrasound scanlines, producing artefacts in the ultrasound image (Scobbie, Punnoose, & Khattab, 2013). We tend to observe a bright white region above where the tongue tip is expected (Mielke et al., 2016) and a discontinuity in the tongue contour where the tongue tip is curled up (Bakst, 2016).

In order to facilitate the task of classifying tongue configurations, the decision tree presented in Figure 4.10 was produced and used throughout the classification process. Tongue shapes were classified three times throughout the course of one year to ensure accuracy. Although discrepancies in the three classifications were rare, such cases were reexamined and the most common configuration of the three was selected.

If we employ the traditional retroflex-bunched classification, the Mid Bunched and Front Bunched configurations have a low tongue tip and the primary constriction is located between the front to mid tongue body (Lawson et al., 2011), so we can consider them to be bunched.

[5]The first four configurations (MB, FB, FU, TU) are identical to those described in Lawson et al. (2011).

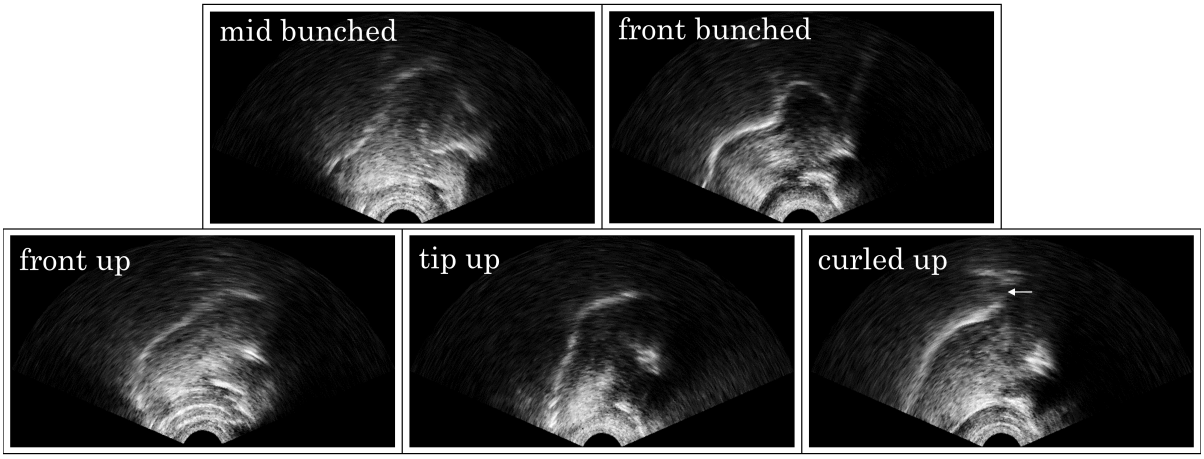



**Figure 4.9:** *Raw ultrasound frames presenting typical examples of each of the five tongue configurations observed in Anglo-English* /r/*. The tongue tip is on the right. The top two images are bunched, while the bottom three are retroflex. The final retroflex configuration exhibits curling up of the tongue which is evident from the bright white line where the tongue tip is expected towards the palate, and a discontinuity in the tongue contour where the tip starts to curl up (indicated with an arrow).*

If we adopt Hamann (2003)'s definition, any sound articulated with the tongue tip behind the alveolar region and involving a displacement of the tongue back towards the pharynx or velum may be considered retroflex. As bunched /r/ has also been shown to include tongue root retraction (Delattre & Freeman, 1968; Proctor et al., 2019) and the drawing inwards of the tongue body away from the lips (Alwan et al., 1997), the main criterion we considered to define retroflexion for /r/ is the raising of the tongue tip, which results in the addition of a sublingual space. The tongue tip and/or tongue front are raised towards the post-alveolar region in the last three configurations of our classification (FU, TU, CU), and so, we therefore consider them to be retroflex. As discussed in Section 2.4, the status of bladal configurations such as the one described in our Front Up category, has been disputed with some researchers who consider them to be bunched rather than retroflex. Although in some raw ultrasound images in our dataset the primary constriction (i.e., the highest point of the tongue) in some Front Up configurations may appear to be the tongue dorsum (as in the Front Up image presented in Figure 4.9), when the corresponding spline is rotated to the occlusal plane, the tongue tip does generally appear to be the primary constriction, or at least pointing up, an example of which

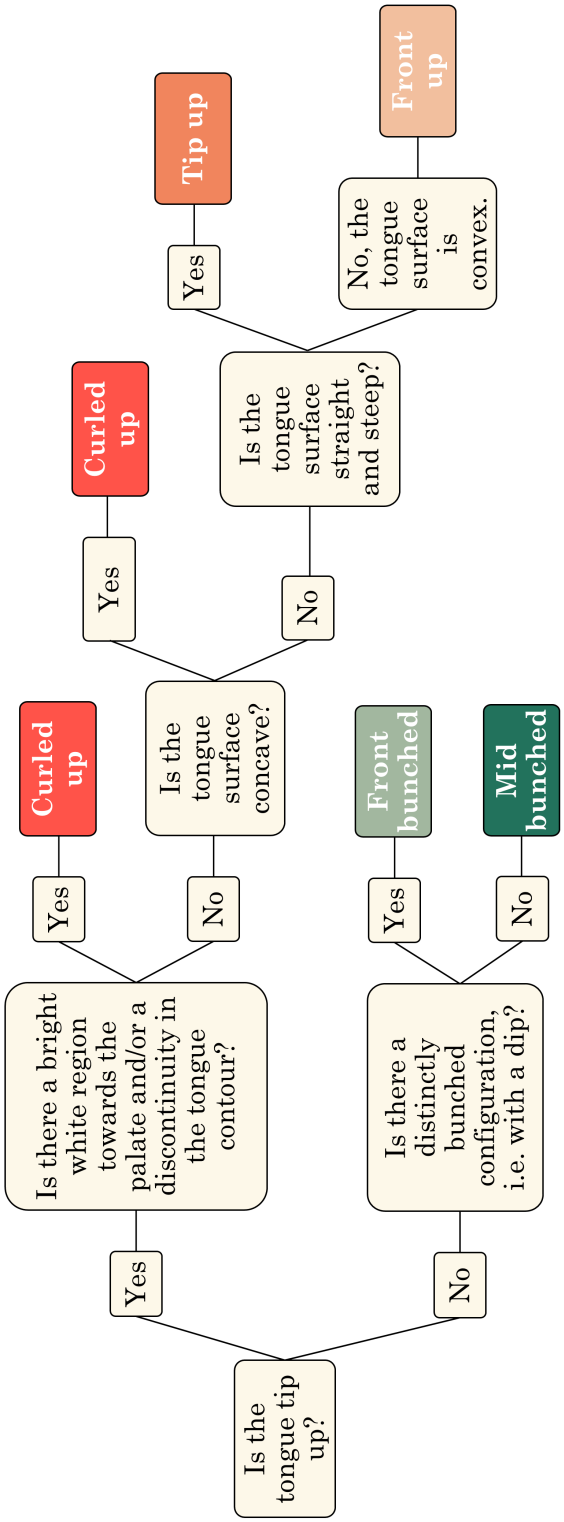



**Figure 4.10:** *Decision tree used to classify tongue configurations for* /ɹ/ *into five distinct categories from ultrasound data.*

can be observed in Figure 4.8. The position of the tongue tip is a further indication that the Front Up configuration exhibits apicality, which is suggestive of retroflexion.

Our classification would place the variant with the highest, most curled up tongue tip, the Curled Up configuration, at one end of the continuum. Curled Up is followed by the Tip Up and Front Up variants respectively. Deciphering which tongue shape is the most bunched category between Mid Bunched and Front Bunched is less evident. Although visualising the tongue contour tracings in speakers who present both configurations revealed that the tongue tip is generally lower in the Mid Bunched than the Front Bunched configuration, the Front Bunched category presents the most obvious bunching of the tongue i.e., with a dip in the tongue surface, or sulcalization (as can be seen in Figure 4.9). Furthermore, the very tip of the tongue is not always visible from ultrasound images and so we err on the side of caution regarding the accuracy of tongue tip tracings. It is hoped that results from this study may provide further insights into which bunched configuration is the most extreme of the two.

### *4.2.7 Measuring lip protrusion*

A profile lip camera was mounted on a bracket attached to the right side of the stabilisation headset at a fixed distance from the mid-line of the speaker's head (Lawson et al., 2019). This profile camera allowed us to film the front-back position of the lips, which we equate to lip protrusion. Quantitative measures of lip protrusion were made using AAA. One image corresponding to a neutral lip configuration (with the lips closed) prior to speech was visually selected per speaker. The image corresponding to maximum lip protrusion was visually identified for each production of /r/ by holistically examining sequential video frames. Lip protrusion was measured by calculating the difference between maximum protrusion and the speaker's neutral lip protrusion setting. To obtain quantitative measurements, a fiducial line[6] was positioned to intersect the lip corner during each speaker's neutral lip image. This fiducial had previously been scaled (in centimetres) to a physical ruler positioned along the mid-line of the stabilisation headset (Lawson et al., 2019), and ran parallel to the upper and

[6]We define fiducial as a fixed line used as a basis of reference and measure.

lower edges of the video pane. Each speaker was assigned one lip corner fiducial which was used for all his/her protrusion measures. For the same neutral lip image, a line was positioned to touch the lower and upper lip edge, intersecting the neutral lip corner fiducial. Using AAA, we calculated the distance from the origin of the fiducial to where the lip edge line crossed, yielding a value (in centimetres) for the neutral lip position. As depicted in Figure 4.11, the neutral lip distance measurement (distance 1) was subtracted from the maximum protrusion distance for /r/ (distance 2) yielding final lip protrusion values.

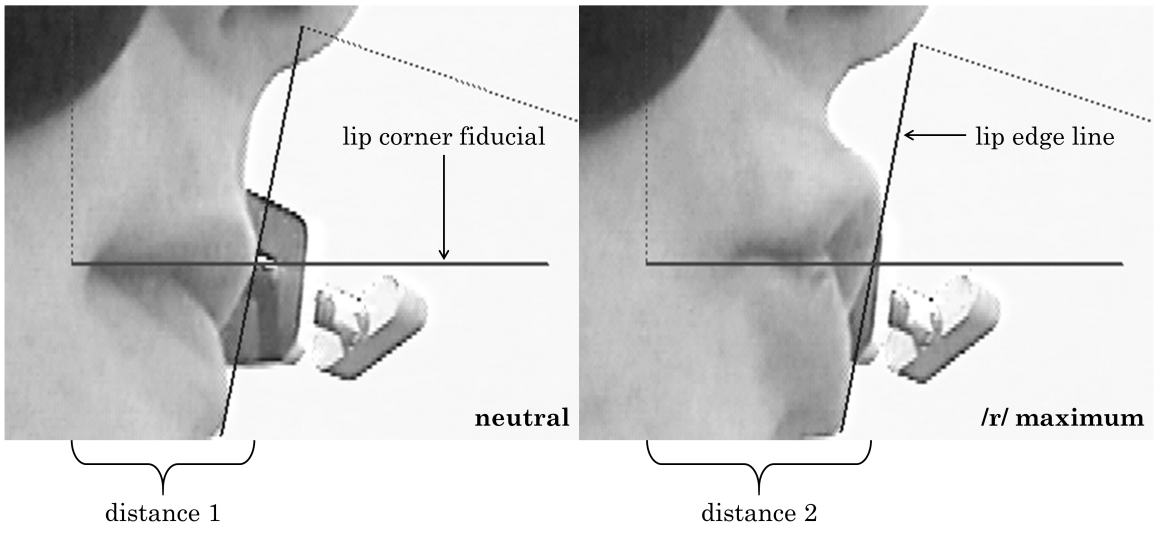



**Figure 4.11:** *Lip protrusion measure. Distance 1 is subtracted from distance 2.*

### 4.2.8 Statistical analysis

Statistical analysis was implemented in R (R Core Team, 2018) using the lmer() function of the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) to perform a series of linear mixed-effects models. We tested the significance of main effects to model fit using likelihood ratio tests with the mixed() function in the *afex* package (Singmann, Bolker, Westfall, & Aust, 2015). Model residuals were plotted to test for deviations from homoscedasticity or normality. The lmerTest library (Kuznetsova, Brockhoff, & Christensen, 2017) was used to calculate indications

of significance within the final models, which uses values derived from Satterthwaite (1946)'s approximations for the degrees of freedom. The resulting *p*-values are provided in the model summary tables. R syntax for each model is presented underneath each model summary table. Plots of the predicted effects from final models were generated with the sjPlot package (Lüdecke, 2018).

## 4.3 RESULTS

### *4.3.1 Classification of tongue shapes*

One word for each of the eight lexical set vowels was selected from non-hyperarticulated productions of word-initial /r/ in order to classify tongue shapes (*reed, red, room, ring, rack, run, raw, rot*). All subjects had an observable tongue body gesture for /r/. Visual classification of tongue configurations yielded the results presented in Table 4.3. Out of the 24 speakers, 7 produced only bunched /r/ configurations, 14 produced only retroflex, and 3 used both. Our data therefore contradict traditional descriptions of Anglo-English /r/ in that speakers do not only produce /r/ with a tip up articulation. However, we observed double the number of speakers producing only retroflex compared to speakers producing only bunched /r/.

In order to discern any patterns regarding the geographical origin of speakers and their tongue configuration for /r/, the map presented in Figure 4.12 was produced. To make any real claims concerning the relationship between tongue shape and speaker origin, we would require more regionally-stratified data. However, from the present dataset, we note that two subjects (08 & 21) who come from the same town in the North West, Chester, use bunched and retroflex /r/ respectively. The only discernible pattern in our data concerns the subjects who use both retroflex and bunched /r/, as all three come from the South East, although other speakers from the same region were observed using either retroflex or bunched shapes. It is interesting to note that labiodental variants have been established as an accent feature of non-standard accents from the same region (Foulkes & Docherty, 2000).

If we take a more detailed look at tongue configuration going beyond the simplistic retroflex-

| Subject code | Age | Sex | /r/ coding | Shape |
|---|---|---|---|---|
| 05 | 22 | F | MB | bunched |
| 08 | 26 | F | MB | |
| 17 | 27 | F | MB | |
| 10 | 44 | M | FB MB | |
| 03 | 22 | F | FB | |
| 11 | 29 | F | FB | |
| 22 | 23 | F | FB | |
| 29 | 18 | F | MB FB FU TU CU | bunched & retroflex |
| 14 | 23 | F | MB FB CU | |
| 18 | 23 | F | FB FU CU | |
| 02 | 22 | F | FU | retroflex |
| 23 | 33 | F | FU TU | |
| 16 | 25 | F | FU TU | |
| 13 | 54 | F | TU | |
| 12 | 20 | F | FU CU | |
| 15 | 25 | F | FU TU CU | |
| 19 | 28 | F | FU TU CU | |
| 27 | 37 | F | FU TU CU | |
| 28 | 29 | F | FU TU CU | |
| 07 | 22 | F | TU CU | |
| 09 | 21 | F | TU CU | |
| 21 | 41 | F | TU CU | |
| 25 | 55 | F | TU CU | |
| 04 | 53 | M | CU | |

**Table 4.3:** *Observed tongue configurations in 24 subjects divided into three categories ordered from most bunched to most retroflex.*

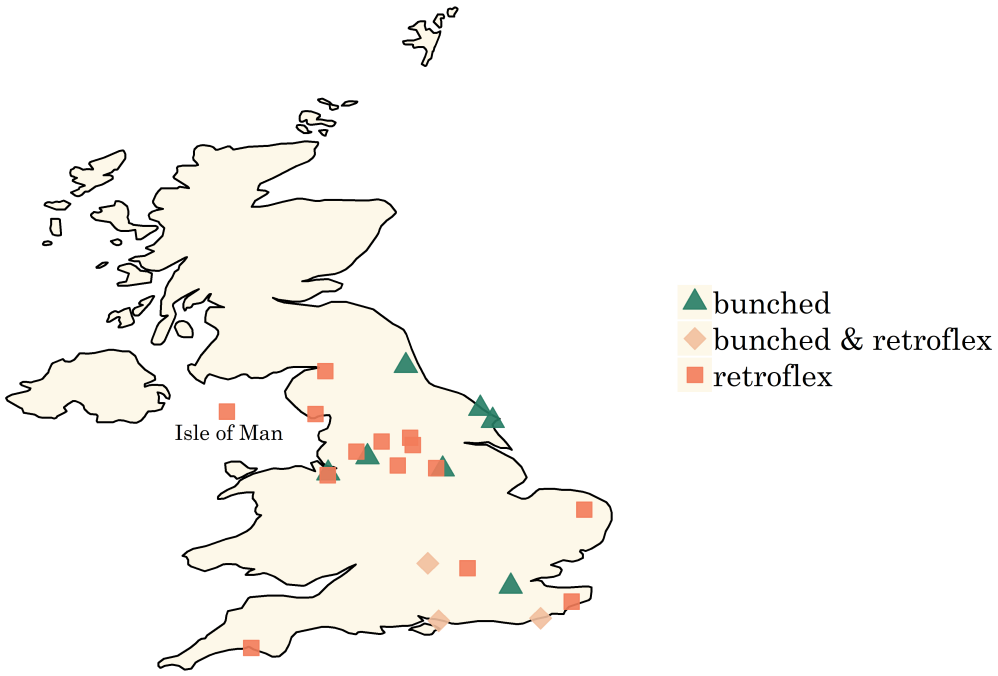



**Figure 4.12:** *Map of speaker origin as a function of tongue configuration for* /r/.

bunched distinction, based on our classification using five distinct shapes as presented in Section 4.2.6 (p. 120), we observe 9 out of the 24 subjects using one configuration exclusively in their non-hyperarticulated productions, 6 of which are bunchers. In fact, all bunchers but one use one tongue configuration across all vowel contexts. The remaining 15 speakers use multiple configurations. One buncher (speaker 10) uses the Front Bunched configuration in all vowel contexts except before the FLEECE vowel, where the Mid Bunched shape is used instead. Among the 17 retroflexers in the dataset, 13 of them use the extreme Curled Up configuration at least some of the time, which has previously been associated more with Anglo-English than American English. However, only one speaker (speaker 04) produces this extreme Curled Up variant exclusively, leading us to suspect that the following vowel may have a co-articulatory influence on retroflexion in most speakers, which has also been observed in American English (as discussed in Section 2.4, p. 41).

In order to discern any patterns regarding tongue shape and the following vowel, we first need to establish what constitutes a close-front and a open-back vowel in Anglo-English. If

we agree that F2 is an acoustic correlate of tongue anteriority and F1 of tongue height, vowel plots should give us some indication of the relative frontness and openness of the vowels in the system. First and second formant values were extracted at the midpoint of a steady state of the vowel from the /r/-initial words in Hertz. Formant values were scaled by means of Lobanov normalisation (Lobanov, 1971). Figure 4.13 shows ellipses to one standard deviation from the Lobanov normalised values. One striking observation is the frontness of the GOOSE vowel which is a known feature of UK accents, especially in SSBE and many varieties of Scottish English (e.g., Ferragne & Pellegrino, 2010; Harrington et al., 2011; Lawson et al., 2019). In terms of F2, GOOSE is by far the most variable of all the vowels in our dataset, with some tokens approaching the space occupied by FLEECE while others have an F2 closer to that of LOT. As previously discussed, articulatory studies have shown that the GOOSE vowel, while still rounded, can no longer be considered a back vowel in many varieties of English (e.g., Harrington et al., 2011; King & Ferragne, 2018; Lawson et al., 2019). Our formant data indicate that while some productions of the GOOSE vowel are fronted, others remain relatively back. This may be a result of having a large number of subjects from the north of England in our dataset ($n$=16) who have previously been shown to present less /u/-fronting than southerners (Ferragne & Pellegrino, 2010; Lawson et al., 2019). The STRUT vowel is also rather variable with some tokens having much higher F1 values than others, which presumably reflects dialectal differences concerning the FOOT-STRUT split. The backest vowel of the system is THOUGHT and the frontest is FLEECE. If retroflexion is favoured by back rather than front vowels, we would expect *raw* to exhibit more retroflexion than *reed*. However, if retroflexion favours open vowels over close vowels, we would expect /r/ preceding the TRAP vowel in *rack* to induce the most retroflexion, as it is the most open vowel in our dataset.

To examine to what extent the following vowel affects retroflexion, we considered the data from speakers who use at least one of the three retroflex configurations ($n$=17). Exclusively bunched /r/ users ($n$=7) were therefore excluded from this analysis. The proportion of each of the five /r/ configurations was plotted as a function of the following vowel in Figure 4.14. As predicted, the FLEECE vowel has the least retroflexion with less than 6% of the tokens presenting

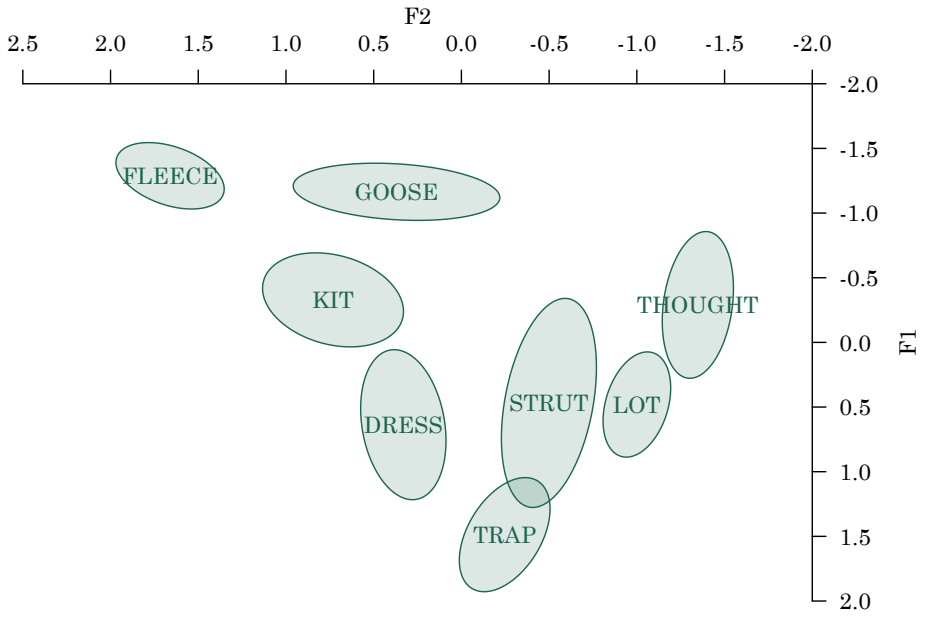



**Figure 4.13:** *Lobanov-transformed vowel plot with one standard-deviation ellipses.*

the extreme Curled Up variant. We observe that in the speakers who use both retroflex and bunched variants, the bunched tokens are only used in /ɹ/ followed by the frontest vowels of the system (FLEECE, GOOSE, KIT, DRESS). It may be that in these speakers, retroflexion is incompatible with front vowels and as a result, bunched configurations are used instead. The most retroflexion was observed preceding the LOT vowel with around 75% of tokens presenting the extreme Curled Up tongue configuration. Our data seem to be consistent with previous work on American English in that retroflexion is favoured by open-back vowels. Although the THOUGHT vowel is the backest vowel of the system, LOT favours retroflexion more, perhaps because it is more open. However, TRAP is more open than STRUT but presents less retroflexion, perhaps because STRUT is generally further back. It seems then that both tongue position and height of the neighbouring vowel affect the tongue configuration used for /ɹ/.

For visualisation purposes, Figure 4.15 presents tongue contour tracings for each speaker's /ɹ/ production at the point of maximal constriction preceding the FLEECE vowel (solid line) and the LOT vowel (dashed line) ordered from most bunched to most retroflex. Asterisks correspond

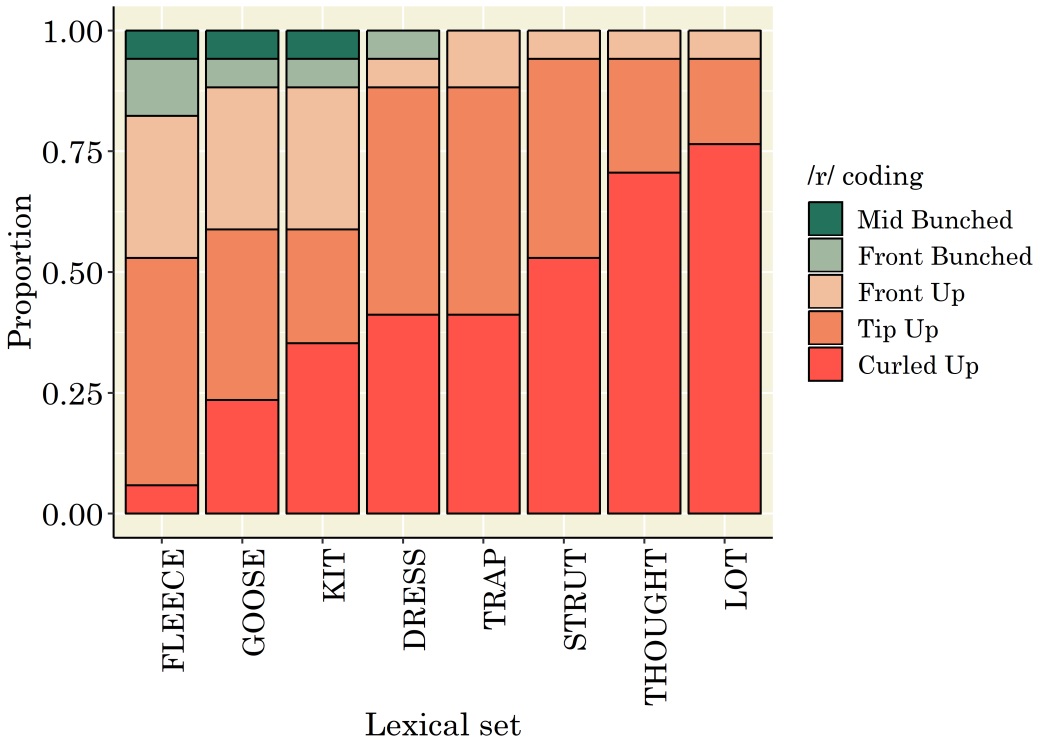



**Figure 4.14:** *Proportion of tongue configurations for /ɹ/ as a function of the following vowel in retroflex users.*

to speakers who were coded as using more than one of the five tongue configurations. Even in speakers who are not considered to present multiple tongue shapes for /r/, we observe differences in tongue position between the two contours. The tongue is generally more anterior preceding FLEECE than it is preceding LOT, which is almost certainly a result of coarticulation. This observation may have an influence on the extent of accompanying lip protrusion. As we have already noted in Chapter 2 (Section 2.8, p. 64), extending the front cavity results in lowering of F3 for /r/. Assuming that the front cavity is smaller for /r/ followed by the FLEECE vowel than it is for /r/ followed by LOT due to coarticulation, in order to maintain a stable acoustic output for /r/ across all vowel contexts, speakers may compensate by using varying amounts of lip protrusion. /r/ followed by the FLEECE vowel may exhibit more protrusion than more open, back vowels, although we do not yet know to what extent the labial properties of neighbouring vowels have a co-articulatory influence on the lips for /r/.

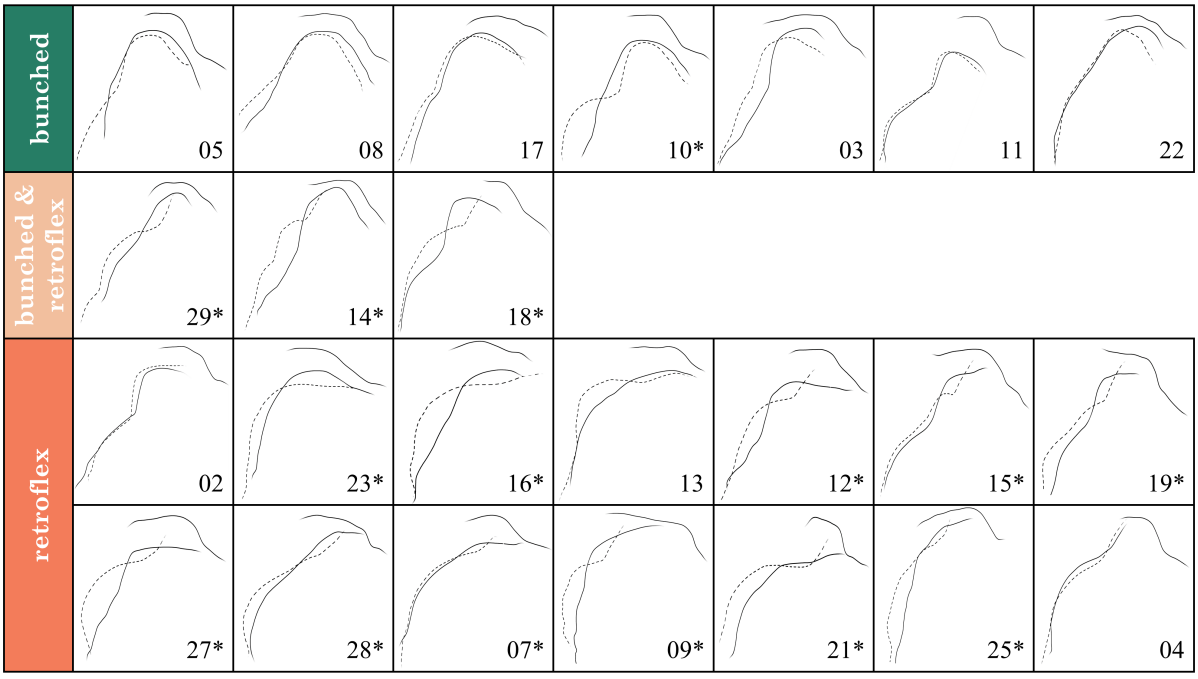



**Figure 4.15:** *Tongue contour tracings ordered from most bunched to most retroflex for /r/ preceding the FLEECE (solid line) and the LOT vowel (dashed line). Speakers who use more than one of the five tongue configurations are indicated with an asterisk. The tongue tip is at the right side of the image. The palate is traced in the top curve for each speaker.*

### 4.3.2 *The influence of tongue shape on lip protrusion*

The influence of tongue shape on lip protrusion was first assessed from the non-hyperarticulated /r/ productions. In the three speakers who produced both retroflex and bunched /r/ configurations, the bunched variants had on average more lip protrusion than retroflex ones, as presented in the plots in Figure 4.16, which include the mean and standard deviation where possible (subject 18 only produced one bunched token). This result therefore suggests that the degree of lip protrusion may be dependent on tongue shape, with bunched tongue shapes exhibiting more accompanying protrusion than retroflex ones.

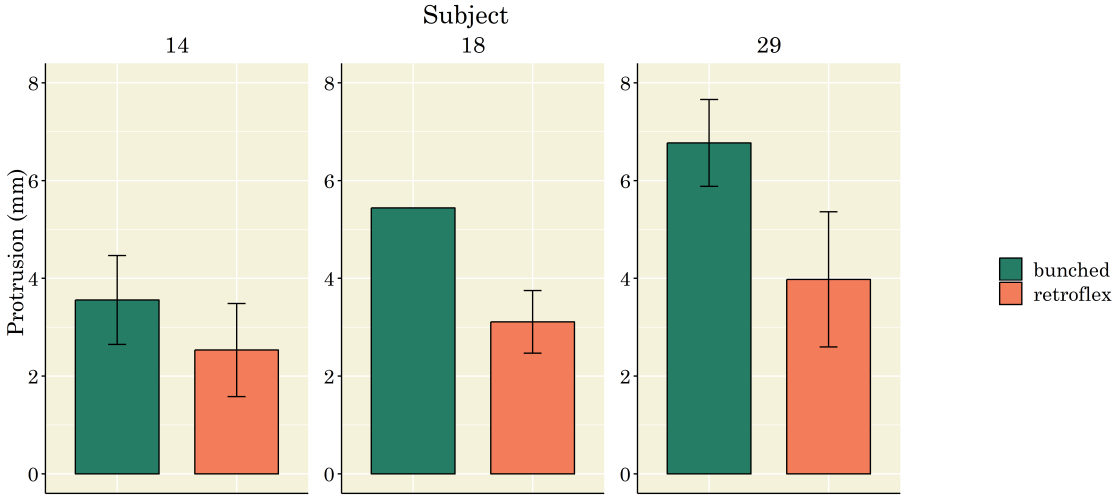



**Figure 4.16:** *Mean and standard deviation lip protrusion values in the three speakers who produce both retroflex and bunched tongue configurations in millimetres.*

In order to assess whether different tongue configurations are accompanied by different degrees of lip protrusion for /r/ in all speakers, a linear mixed-effects regression analysis was performed. The fixed factors were /r/ Coding (CU, TU, FU, FB, MB) and Vowel (FLEECE, GOOSE, KIT, DRESS, TRAP, STRUT, THOUGHT, LOT) and the random structure included by-Speaker random intercepts[7]. There was a statistically significant main effect of both tongue configuration ($\chi^2(4) = 29.74$, $p < 0.001$) and following vowel ($\chi^2(7) = 34.28$, $p < 0.001$) on lip protrusion.

[7]The inclusion of by-item varying intercepts resulted in a singular fit, presumably because, given the limited dataset, the main effect of vowel captures all the item variance.

The final model output is presented in the model summary in Table 4.4.

| Predictor | Estimate | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 2.15 | 0.47 | 4.61 | $< .001^{***}$ |
| /r/ Coding TU | -0.005 | 0.26 | -0.02 | 0.99 |
| /r/ Coding FU | -0.37 | 0.37 | -1.00 | 0.32 |
| /r/ Coding FB | 2.03 | 0.42 | 4.79 | $< .001^{***}$ |
| /r/ Coding MB | 1.40 | 0.56 | 2.51 | $0.02^{*}$ |
| Vowel GOOSE | 0.13 | 0.26 | 0.51 | 0.61 |
| Vowel KIT | -0.67 | 0.26 | -2.60 | $0.01^{*}$ |
| Vowel DRESS | -0.74 | 0.27 | -2.75 | $0.01^{*}$ |
| Vowel TRAP | -0.48 | 0.27 | -1.80 | 0.08 |
| Vowel STRUT | -0.11 | 0.27 | -0.39 | 0.70 |
| Vowel THOUGHT | 0.15 | 0.28 | 0.55 | 0.59 |
| Vowel LOT | 0.38 | 0.29 | 1.32 | 0.19 |

*Protrusion* ~ *rCoding* + *Vowel* + (1|*Speaker*)

**Table 4.4:** *Output of a linear-mixed effects regression model predicting lip protrusion. The intercept corresponds to a CU tongue configuration preceding the FLEECE vowel.*

As Table 4.4 indicates, the bunched tongue configurations (FB and MB) are predicted to have significantly more lip protrusion than the extreme Curled Up retroflex. Although more lip protrusion is predicted in FB, by changing the reference level to FB and rerunning the model, we found no significant difference between FB and MB. There was no significant difference between the Curled Up retroflex and the other two retroflex configurations (TU & FU). Figure 4.17 presents the predicted effects of tongue configuration for /r/ on lip protrusion. We observe that the three retroflex configurations pattern together with the least protrusion, as do the two remaining bunched ones, with the most protrusion. As discussed in Section 4.3.1, the Front Up configuration seems to lie somewhere in the middle of the retroflex-bunched continuum with regards to its lingual characteristics. However, we notice that with regards to lip protrusion, Front Up strongly patterns with the Curled Up and Tip Up retroflex configurations. This result further justifies our decision to consider the Front Up configuration a retroflex and not a bunched shape.

With regards to the effect of the following vowel on lip protrusion for /r/, the model predicts

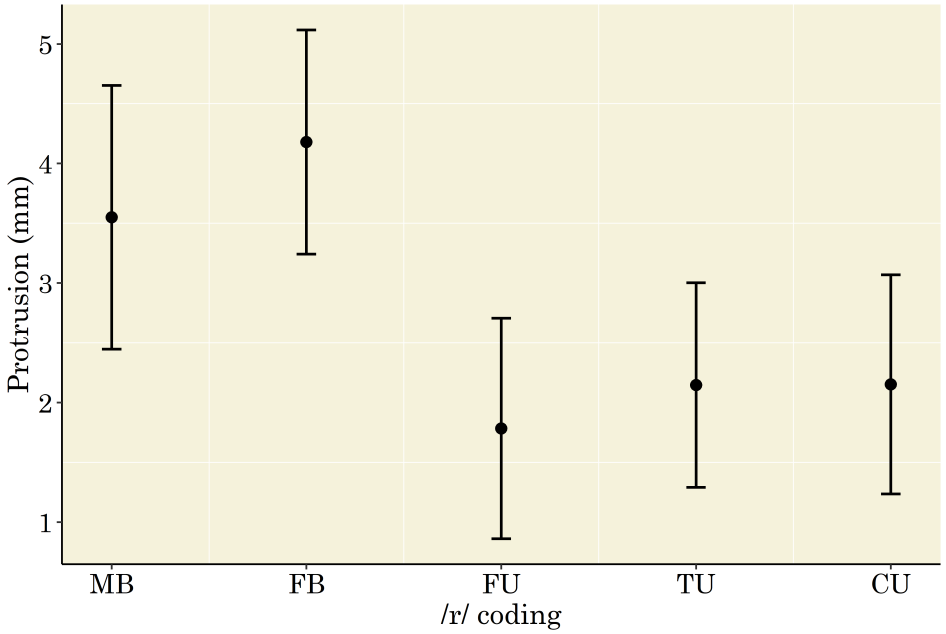



**Figure 4.17:** *Predicted effects of tongue configuration on lip protrusion from a linear-mixed effects regression model. Error bars (here and in all subsequent predicted effects plots) are 95% confidence intervals.*

that the KIT and DRESS vowels have significantly less protrusion than the FLEECE vowel. No significant difference is predicted between the FLEECE vowel and the remaining vowels in the dataset (GOOSE, TRAP, STRUT, THOUGHT, LOT). Figure 4.18 presents the predicted effects of the following vowel on protrusion in /r/ from the model. The model output in Table 4.4 indicates that in the context of the rounded vowels LOT, THOUGHT and GOOSE, /r/ is estimated to have the highest degrees of lip protrusion, suggesting a co-articulatory influence of lip rounding from the following vowel.

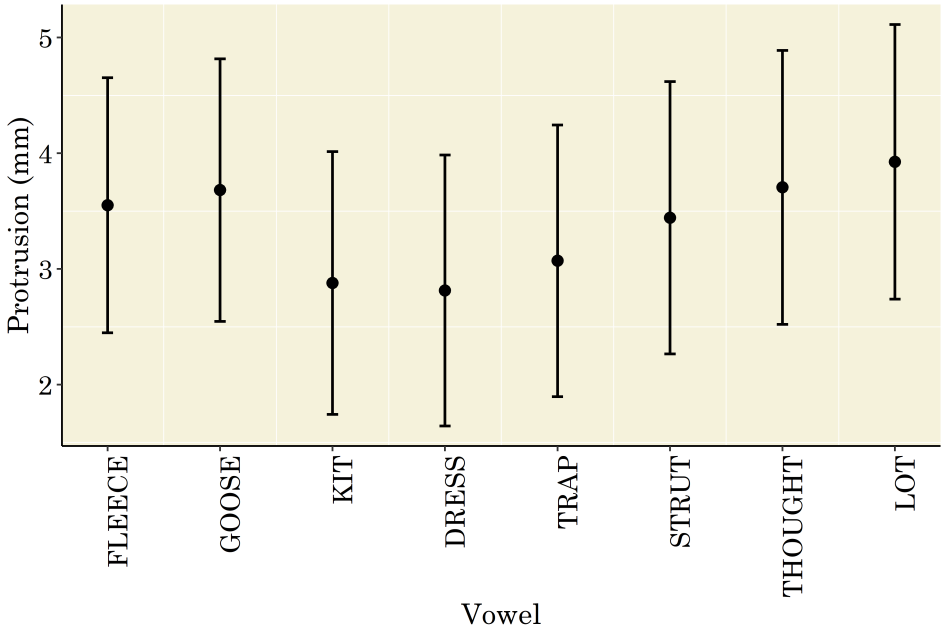



**Figure 4.18:** *Predicted effects of following vowel on lip protrusion from a linear-mixed effects regression model.*

### *4.3.3 /r/ acoustics*

So far, our articulatory analysis has shown that Anglo-English /r/ is produced with a range of tongue shapes ranging from retroflex to bunched. The following vowel has a co-articulatory effect on tongue shape, tongue position and the degree of accompanying lip protrusion. Close-front vowels are produced with less retroflexion than open-back ones. /r/ followed by the front FLEECE vowel is generally produced with a more anterior palatal constriction than /r/ followed by the back LOT vowel. More lip protrusion is observed for /r/ in the context of rounded vowels than non-rounded ones. However, no significant difference in lip protrusion for /r/ was observed between rounded vowels and the FLEECE vowel, which suggests that speakers may compensate for the fronted tongue position in the context of FLEECE by extending the front cavity with lip protrusion. Our analysis further points to the use of lip protrusion as an articulatory strategy used to lengthen the front cavity because bunched tongue shapes, which are produced with little space underneath the tongue, present more lip protrusion than

retroflex ones.

But how do these articulatory configurations affect /r/ acoustics? We will specifically consider the effect of the shape of the tongue for /r/ and the following vowel. Previous research on rhotic Englishes has not found significant differences in the formant frequencies up to F3 between the different possible tongue configurations for /r/. As our dataset contains limited data from male subjects ($n$=2) and as it is well established that speaker sex influences formant values, we will only consider data from the remaining female subjects ($n$=22) in our acoustic analysis. Firstly, across all non-hyperarticulated productions of /r/ in women, the following mean formant values and their standard deviations (in Hz) were observed:

F1: 421 ± 65

F2: 1 236 ± 224

F3: 1 881 ± 198

Mean formant values are consistent with the range of values observed in previous studies on /r/ in American English (as presented in Chapter 2, Section 2.8, p. 64). Table 4.5 shows mean formant values (in Hz) and their standard deviations according to tongue shape. Unlike previous research on rhotic Englishes, the mean formant values in our dataset do suggest that there may be differences across tongue shapes in formant values, notably with regards to FB, which has a lower mean F3 than the other four shapes.

| /r/ coding | F1 | F2 | F3 |
|---|---|---|---|
| CU | 435 (71) | 1 158 (212) | 1 851 (184) |
| TU | 419 (71) | 1 253 (247) | 1 914 (186) |
| FU | 442 (66) | 1 318 (209) | 1 960 (217) |
| FB | 399 (46) | 1 254 (227) | 1 761 (184) |
| MB | 411 (54) | 1 279 (147) | 2 026 (116) |

**Table 4.5:** *Mean formant values (in Hz) and their standard deviations (in parentheses) for all tongue configurations from most retroflex to most bunched in women.*

Our analysis of articulatory data indicated that the following vowel has a co-articulatory

influence on the production of /r/, which may therefore have acoustic consequences. The UTI data suggested that /r/ in the context of front vowels is generally produced with a fronter lingual constriction than /r/ in the context of back vowels. The following vowel was also a significant predictor of lip protrusion and the highest degrees of lip protrusion seem to occur in the context of rounded vowels. Although it would be tempting to predict that /r/ in the presence of rounded vowels will result in lower formant frequencies than non-rounded vowels, we know from our review of the literature in Chapter 3 that the labial and lingual constrictions work in harmony to shape formant frequencies. Teasing apart the relative impact of the lips and tongue on the acoustics of /r/ is therefore not an easy task and is made more challenging by the fact that in the present dataset, place of articulation and rounding in vowels are confounded, i.e., all back vowels are rounded.

However, we will make a few tentative predictions concerning the effect of the following vowel on the formant frequencies of /r/, focusing on F2 and F3. As we expect F3 to be related to the size of the front cavity, with larger front cavities resulting in lower F3 values, we expect /r/ in the context of the backest vowels of the dataset, THOUGHT and LOT, to result in the lowest F3 frequencies. Incidentally, these two vowels induce the highest degrees of lip protrusion in /r/, probably because they are produced with lip rounding. As lip protrusion extends the front cavity, these vowels should have a further lowering effect on F3 for /r/ (although the absence of a non-rounded equally back vowel prevents us from testing this claim in the present dataset). The vowel plot in Figure 4.13 indicated that the back vowels THOUGHT, LOT and STRUT had the lowest F2 values. As we know that F2 is particularly impacted by lip rounding in back vowels (cf. Chapter 3), we predict that /r/ will be produced with the lowest F2 frequencies when followed by the back rounded vowels THOUGHT and LOT. Therefore, in the context of the THOUGHT and LOT vowels, both F3 and F2 should be at their lowest for /r/. We note that we chose to follow previous studies on English /r/ and not consider F1 in our analysis. We will therefore present separate statistical analyses for F3 and F2. The absolute height of F2 and F3 were considered rather than the relative distance between them because absolute values allowed us to make clear predictions regarding the co-articulatory effect of the following vowel.

### F3

Figure 4.19 presents box plots of the raw F3 values (in Hz) for each of the five tongue configurations, which like Table 4.5, again suggests that FB has a lower F3 than the other configurations. The median value of FB is lower than all the other tongue configurations and although the interquartile range is small, FB has the most outliers.

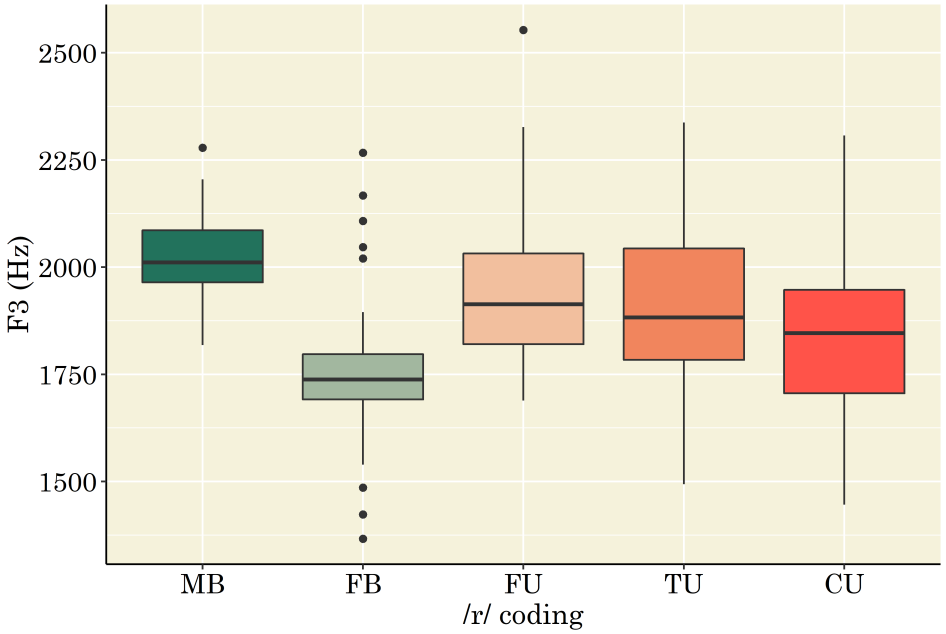



**Figure 4.19:** *Box plots of raw F3 values (in Hz) for each of the five tongue configurations for* /r/. *The boxes (here and in all subsequent box plots) represent the interquartile range, i.e., the middle 50% of values. Whiskers extend to the highest and lowest values, excluding outliers (in circles). An outlier is any data value that lies more than one and a half times the interquartile range outside the box. A line across the box indicates the median.*

Table 4.6 shows mean and standard deviation F3 values of /r/ (in Hz) productions in women according to the following vowel, which indicates that /r/ following the close-front FLEECE vowel results in the highest F3 value on average. While /r/ following the lowest, backest vowel of the system (LOT) has the lowest F3 value on average. These results suggest that the more open and more back the following vowel, the lower the F3, which is also apparent from the box plots presented in Figure 4.20.

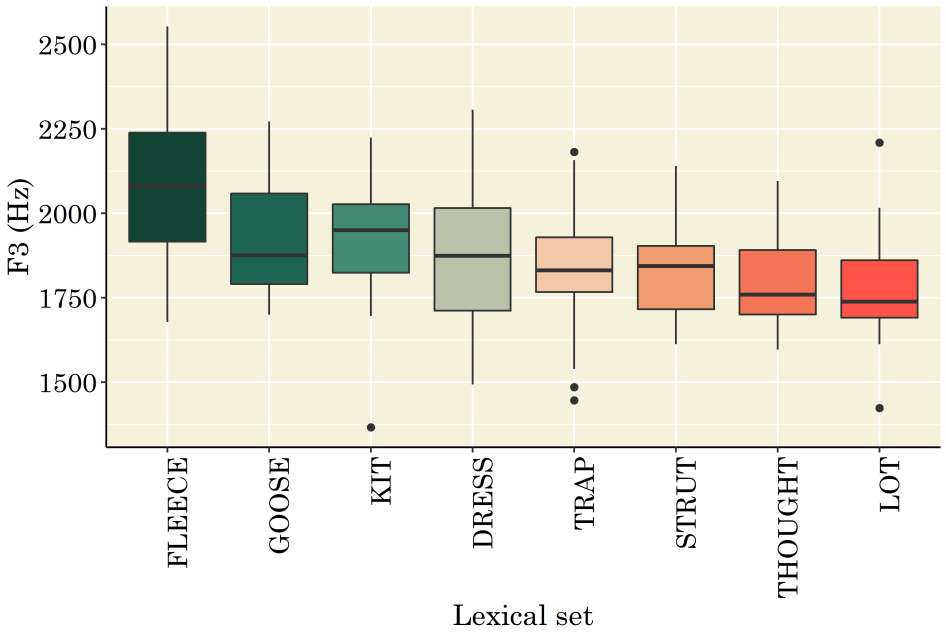



**Figure 4.20:** *Box plots of raw F3 values (in Hz) ordered from highest to lowest according to the following vowel.*

| Lexical set | F3 |
|---|---|
| FLEECE | 2 079 (224) |
| GOOSE | 1 925 (163) |
| KIT | 1 918 (191) |
| DRESS | 1 889 (197) |
| STRUT | 1 834 (143) |
| TRAP | 1 823 (192) |
| THOUGHT | 1 797 (145) |
| LOT | 1 785 (170) |

**Table 4.6:** *Mean F3 values (in Hz) and their standard deviations (in parentheses) for* /ɪ/ *according to the following vowel ordered from highest to lowest.*

To test whether there are statistically significant differences in F3 for /ɹ/ between the different tongue configurations and the following vowel, we performed a linear mixed-effects analysis. The fixed factors were /ɹ/ Coding (CU, TU, FU, FB, MB) and Vowel (FLEECE, GOOSE, KIT, DRESS, TRAP, STRUT, THOUGHT, LOT) and the random structure included by-Speaker random intercepts. Likelihood ratio tests revealed that there was a statistically significant effect of the following vowel on F3 ($\chi^2(7) = 52.13$, $p < 0.001$) but not of tongue configuration ($\chi^2(4) = 4.32$, $p = 0.36$). The final model output is presented in the model summary in Table 4.7. The model predicts all vowels to have a significantly lower F3 than the FLEECE vowel. According to the model, the lowest F3 values occur in /ɹ/ followed by the back THOUGHT and LOT vowels, following our prediction. Furthermore, these results are in line with previous work on English /ɹ/ because tongue configuration was not a statistically significant factor, contrary to what the mean raw values would indicate. When individual variation is taken into account, any apparent differences in F3 between tongue configurations disappear. Indeed, the model's marginal $R^2$, which is the variance described only by the main effects is 25.03%. The conditional $R^2$, which is the variance described by the main *and* the random effects is much higher at 61.48%.[8] The model also predicts speaker intercepts to range from 1 838 to 2 294 Hz.

[8]Conditional and marginal $R^2$ were calculated using the r.squaredGLMM() function in the MuMIn package (Barton, 2018).

| Predictor | Estimate | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 2 037.16 | 49.68 | 41.01 | $< .001^{***}$ |
| /r/ Coding TU | 19.83 | 34.19 | 0.58 | 0.57 |
| /r/ Coding FU | 62.99 | 49.56 | 1.27 | 0.21 |
| /r/ Coding FB | 9.22 | 52.91 | 0.17 | 0.87 |
| /r/ Coding MB | 128.32 | 69.55 | 1.84 | 0.07 |
| Vowel GOOSE | -151.02 | 36.89 | -4.09 | |
| Vowel KIT | -156.63 | 37.35 | -4.19 | |
| Vowel DRESS | -172.31 | 38.36 | -4.49 | |
| Vowel TRAP | -240.69 | 38.22 | -6.30 | $< .001^{***}$ |
| Vowel STRUT | -226.11 | 39.14 | -5.78 | |
| Vowel THOUGHT | -259.69 | 40.45 | -6.42 | |
| Vowel LOT | -271.39 | 40.99 | -6.62 | |

*F3 ~ rCoding + Vowel + (1|Speaker)*

**Table 4.7:** *Output of a linear-mixed effects regression model predicting F3. The intercept corresponds to a CU tongue configuration preceding the FLEECE vowel.*

### F2

Figure 4.21 presents box plots of the raw F2 values (in Hz) for each of the five tongue configurations. F2 appears to be lowest in the most extreme retroflex CU configuration, although variability across the configurations is evident. Table 4.8 shows mean and standard deviation F2 values (in Hz) for /r/ productions in women according to the following vowel. As predicted, in the context of the back rounded THOUGHT and LOT vowels, /r/ has the lowest F2 on average. The box plots in Figure 4.22 paint a similar picture.

To test whether there are statistically significant differences in F2 for /r/ between the different tongue configurations and the following vowel, we performed a linear mixed-effects analysis in the same manner as the previous regression analysis for F3. The fixed factors were /r/ Coding (CU, TU, FU, FB, MB) and Vowel (FLEECE, KIT, DRESS, TRAP, GOOSE, STRUT, LOT, THOUGHT) and the random structure included by-Speaker random intercepts. Like the F3 regression model, likelihood ratio tests revealed that there was a statistically significant effect of the following vowel on F2 ($\chi^2(7) = 54.08$, $p < 0.001$) but not of tongue configuration ($\chi^2(4) = 2.11$, $p = 0.71$). The final model output is presented in the model summary in Table 4.9.

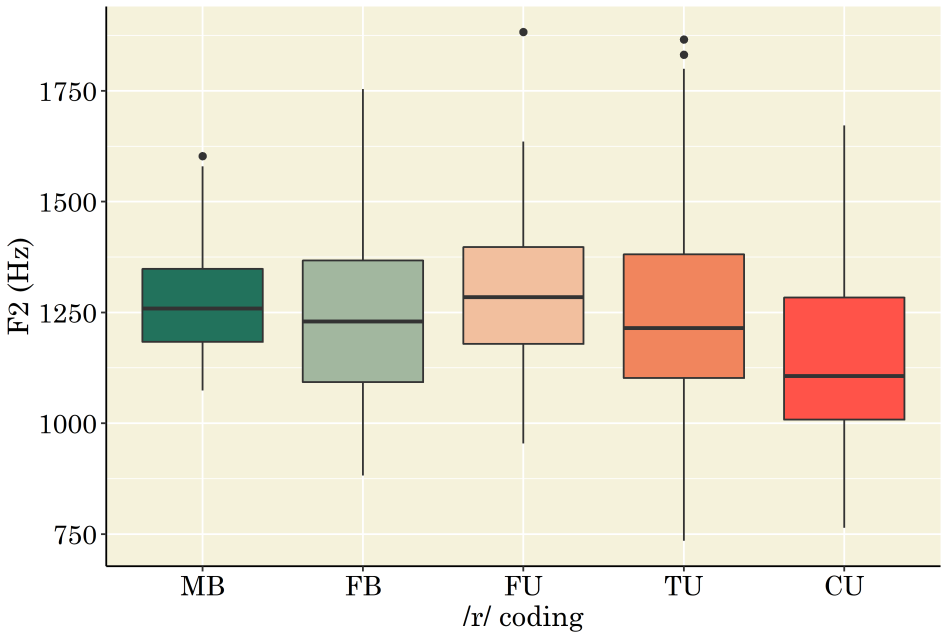



**Figure 4.21:** *Box plots of raw F2 values (in Hz) for each of the five tongue configurations for* /ɹ/.

| Lexical set | F2 |
|---|---|
| FLEECE | 1 436 (272) |
| KIT | 1 294 (219) |
| DRESS | 1 289 (210) |
| TRAP | 1 278 (175) |
| GOOSE | 1 269 (186) |
| STRUT | 1 182 (151) |
| LOT | 1 094 (173) |
| THOUGHT | 1 048 (141) |

**Table 4.8:** *Mean F2 values (in Hz) and their standard deviations (in parentheses) for* /ɹ/ *according to the following vowel ordered from highest to lowest.*

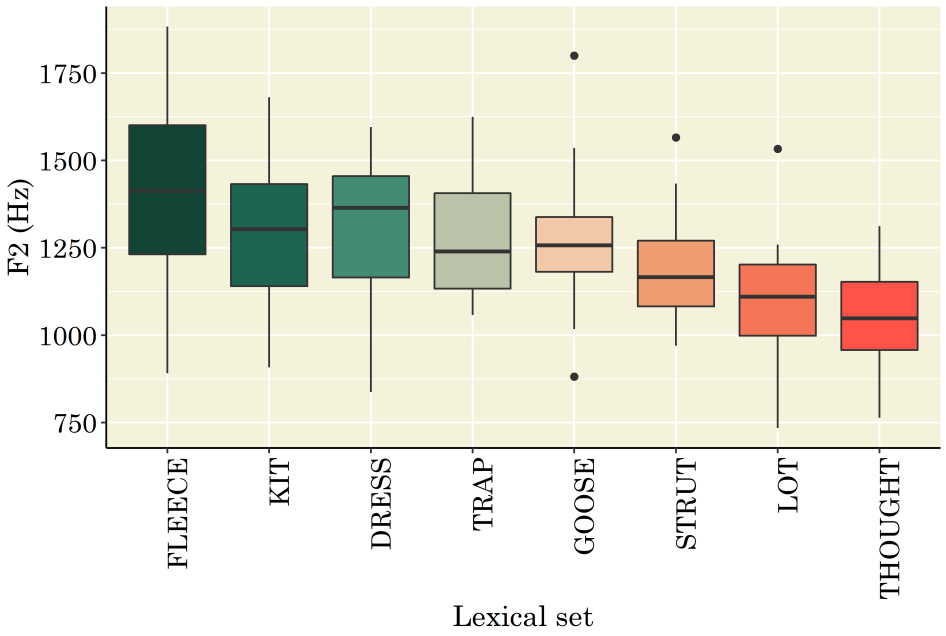



**Figure 4.22:** *Box plots of raw F2 values (in Hz) ordered from highest to lowest according to the following vowel.*

Like in previous studies, no significant difference in F2 is observed between the five tongue configurations. The model output follows our prediction that the lowest F2 for /r/ occurs in the context of the back rounded vowels THOUGHT and LOT. We stress that this analysis does not allow us to assess the relative contributions of the tongue and lips in the following vowel to the frequency of F2 for /r/ due to the absence of a non-rounded back vowel. Although one could be tempted to consider the STRUT vowel, we err on the side of caution given the fact that STRUT may be realised as the rounded [ʊ] vowel in linguistic northerners, who incidentally make up the majority of the dataset ($n$=16). However, both the F2 and F3 analyses suggest a co-articulatory influence of the following vowel on /r/.

| Predictor | Estimate | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 1 406.01 | 57.49 | 24.46 | < .001 |
| /r/ Coding TU | -1.60 | 43.31 | -0.04 | 0.98 |
| /r/ Coding FU | 61.06 | 61.19 | 1.00 | 0.32 |
| /r/ Coding FB | 23.84 | 58.18 | 0.41 | 0.69 |
| /r/ Coding MB | 73.49 | 74.65 | 0.98 | 0.33 |
| Vowel KIT | -141.47 | 50.19 | -2.82 | 0.005** |
| Vowel DRESS | -131.78 | 51.24 | -2.57 | 0.02* |
| Vowel TRAP | -143.90 | 51.02 | -2.82 | 0.005** |
| Vowel GOOSE | -165.52 | 49.65 | -3.33 | 0.001** |
| Vowel STRUT | -237.26 | 52.07 | -4.56 | < .001*** |
| Vowel LOT | -325.53 | 54.23 | -6.00 | < .001*** |
| Vowel THOUGHT | -371.51 | 53.59 | -6.93 | < .001*** |

*F2 ~ rCoding + Vowel* + (1|*Speaker*)

**Table 4.9:** *Output of a linear-mixed effects regression model predicting F2. The intercept corresponds to a CU tongue configuration preceding the FLEECE vowel.*

### 4.3.4 *Hyperarticulated productions of* /r/

In order to compare non-hyperarticulated with hyperarticulated productions of /r/, all /r/ tokens produced after the first recognition error made by the simulated 'silent speech' reader were coded as hyperarticulated. Productions made prior to the initial computer error in the 'silent speech' (non-hyperarticulated) session were therefore not included. All nine /r/-initial words produced in the session in which the computer made no recognition errors were considered to be non-hyperarticulated. For statistical analysis, the dichotomous tongue shapes for /r/ (i.e., bunched and retroflex) will be considered rather than the five configurations to increase experimental power.

#### Modifications to tongue shape

To assess changes in tongue shape from non-hyperarticulated to hyperarticulated /r/ productions, the five tongue configurations were transformed into a numeric scale from zero to four with zero being the most bunched (Mid Bunched) and four being the most retroflex

(Curled Up). The mean tongue shape was then calculated for each speaker according to context (non-hyperarticulated and hyperarticulated). The resulting means were then transformed into a percentage by multiplying by 25. We consider this percentage to correspond to a measure of the rate of retroflexion: a speaker who only produces the most extreme Curled Up (CU) shape would obtain a value of 100%, while a speaker who exclusively uses the most bunched Mid Bunched (MB) shape would obtain 0%. As previously discussed, the Front Up (FU) shape is considered to lie in the middle of the retroflex-bunched continuum. As a result, a speaker who obtains 50% retroflexion produces Front Up configurations exclusively. Rate of retroflexion in the hyperarticulation context increased in 10 of the 14 exclusively retroflex users. In the remaining 4 retroflex users, rate of retroflexion remained the same, although one speaker had already obtained a retroflexion rate of 100% in the non-hyperarticulated context. In 5 out of the 7 bunchers, rate of retroflexion did not change in hyperarticulation. In the remaining 2 bunchers, retroflexion decreased, in other words, bunching increased. In the 3 speakers who present both bunched and retroflex tongue shapes, retroflexion increased in one, while bunching increased in 2. Figure 4.23 shows the mean percentage change in retroflexion from non-hyperarticulated to hyperarticulated /r/ productions for each speaker. The colours correspond to the tongue shape or shapes the speakers were coded to use, i.e., retroflex, retroflex and bunched, or bunched. These results indicate that although 9 speakers present no change in tongue shape, the remaining 15 use more 'extreme' tongue shapes in hyperarticulation. Three of the nine speakers who showed no change already produced the most extreme bunched or retroflex tongue shapes in the non-hyperarticulated context across the board. In speakers who showed a change in the hyperarticulated context, exclusively retroflex users produce more retroflexion, exclusive bunchers produce more bunched shapes and speakers who use both retroflex and bunched shapes presented either more retroflex or more bunched shapes.

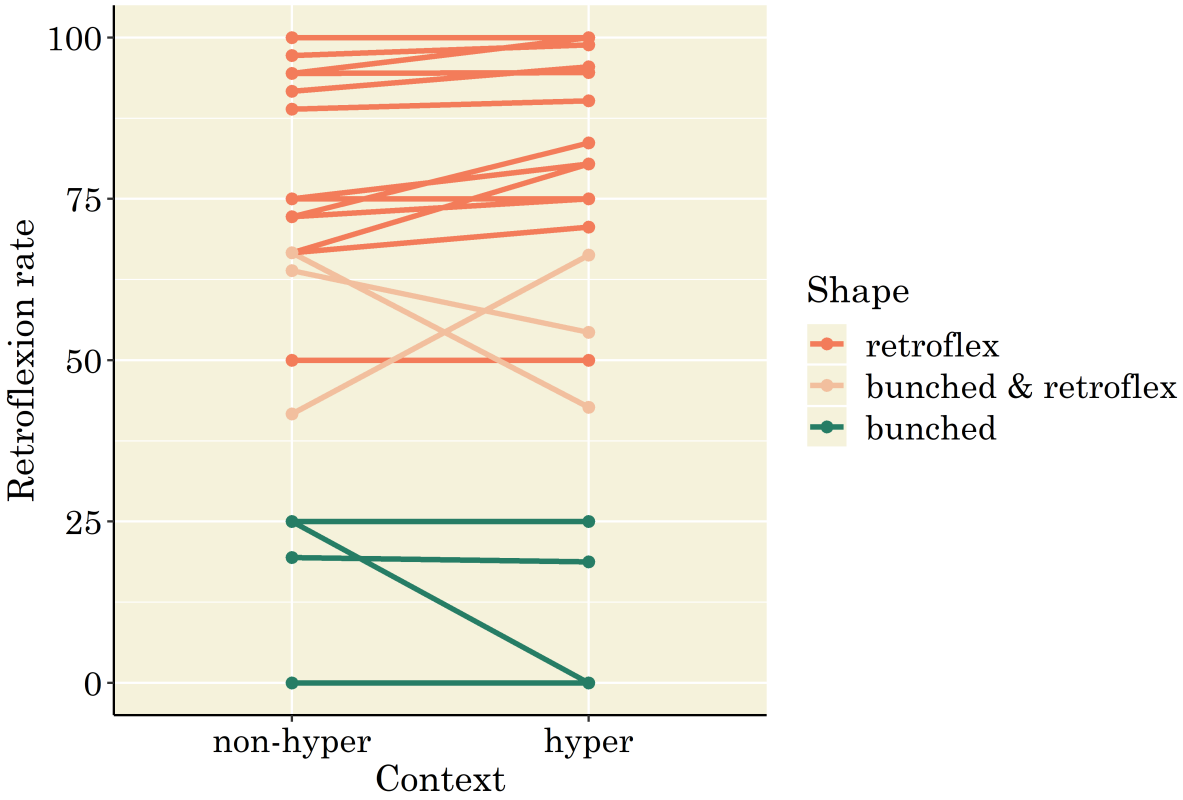



**Figure 4.23:** *Percentage of retroflexion in non-hyperarticulated and hyperarticulated productions of /ɹ/ for each speaker. Colours indicate the shape of the tongue.*

We predicted that a possible hyperarticulation strategy in retroflex users could be the use of more retroflexion, in order to increase the size of the sublingual space and lower F3. Using the same technique as in Section 4.3.1, the proportion of each of the five tongue configurations was plotted as a function of the following vowel for both non-hyperarticulated and hyperarticulated productions of /r/ in speakers who use at least one of the three retroflex configurations ($n$=17). The results are presented in Figure 4.24. Although the left non-hyperarticulated plot is identical to the one presented in Figure 4.14, for ease of comparison, it has been presented again here. We generally observe a higher proportion of the Curled Up (CU) configuration in the hyperarticulated context, although the proportion of CU is smaller in the context of the LOT vowel (76.5% non-hyperarticulated versus 70.6% hyperarticulated). The largest proportional increases in retroflexion occur for /r/ following the vowels TRAP, FLEECE, GOOSE, DRESS. While the latter three were the vowel contexts in which the simulated 'silent speech' programme made recognition errors (i.e., in the words *reed, red, room*), these results indicate that hyperarticulation was generalised to all productions of /r/ even when the computer did not make errors. /r/ followed by the TRAP vowel had the largest proportional increase in extreme retroflexion due to hyperarticulation.

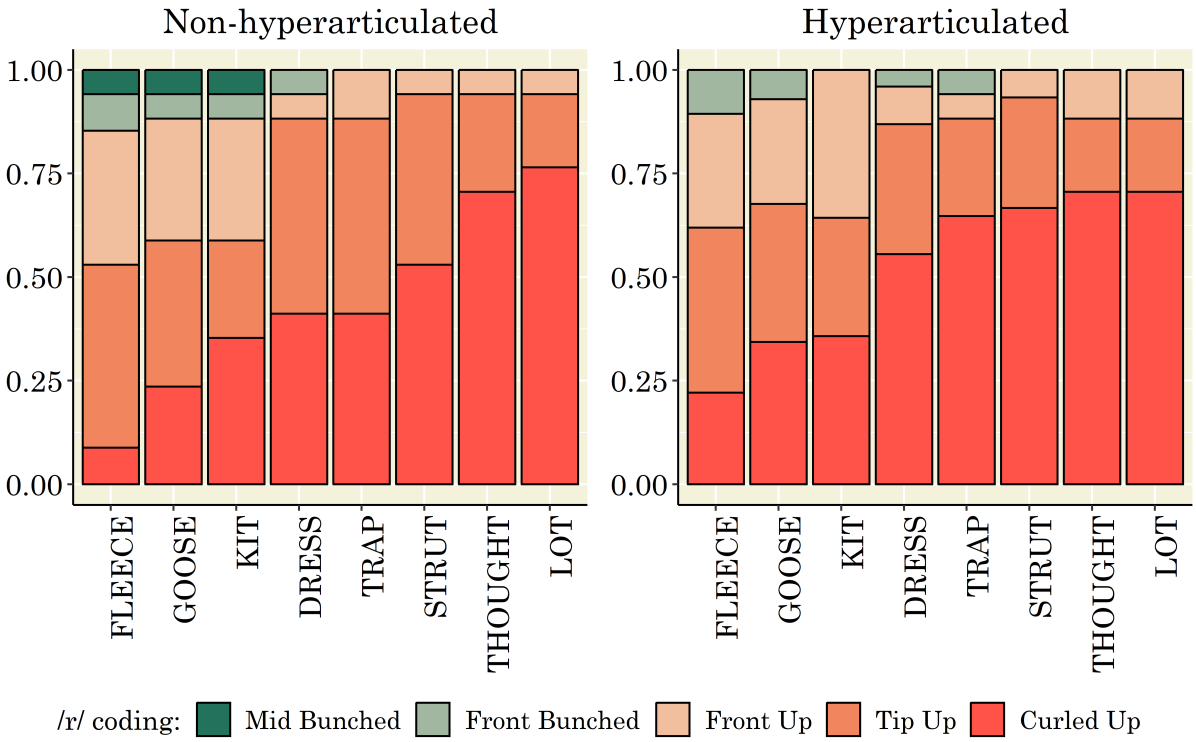



**Figure 4.24:** *Proportion of tongue configurations as a function of the following vowel produced in retroflex users in non-hyperarticulated and hyperarticulated* /ɹ/.

Although tongue shape variation may in some cases be due to coarticulation with the neighbouring vowel, two speakers (14, 18) who use both retroflex and bunched shapes switched from one shape to the other in hyperarticulated /r/ in the same vowel context. Figure 4.25 presents ultrasound images of the word *reed* produced by speaker 18. While in her non-hyperarticulated production, /r/ was produced with a Front Bunched (FB) configuration, /r/ became more retroflex in speech repairs directly following 'w' recognition errors. The first repetition after the misrecognition was produced with a Front Up configuration. She then produced a more extreme Tip Up shape when the computer mistook her /r/ production for 'w' for a second time. Interestingly, she retained her usual FB shape for the /r/ in *reed* when she was presented with 'lead' as the computer's feedback response. We note that as only one repetition was recorded per word in the non-hyperarticulated context, we cannot be sure that she habitually uses a bunched configuration in the context of the FLEECE vowel. We did however record another word containing the same FLEECE vowel, *reap*, which speaker 18 also produced with the same Front Bunched tongue shape.

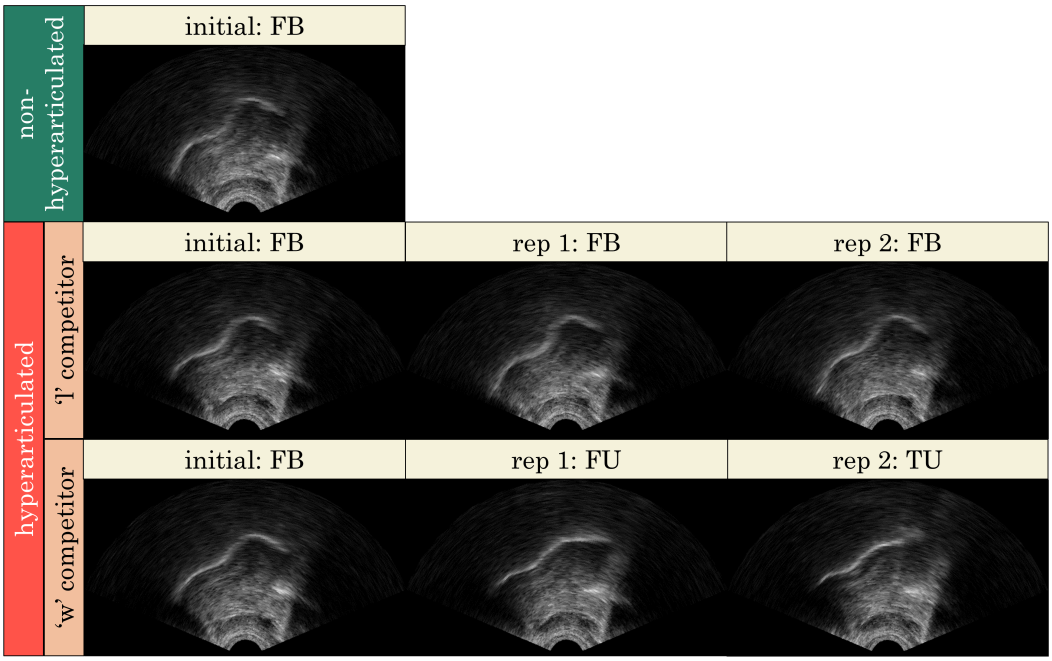



**Figure 4.25:** *Ultrasound tongue images from Speaker 18's productions of the word* reed *which was produced with multiple tongue configurations (FB, FU, TU) with hyperarticulation. The tongue tip is on the right.*

### Lip protrusion

The results from non-hyperarticulated /r/ productions indicated that the degree of lip protrusion may be related to tongue shape with bunched shapes presenting more lip protrusion than retroflex ones. We predicted that hyperarticulated /r/ will be produced with more lip protrusion than non-hyperarticulated /r/ in order to extend the front cavity. While we predicted that retroflexers may further increase the size of the front cavity via more retroflexion, as this strategy is not available to bunchers, bunched /r/ users may present more lip protrusion than retroflexers in the hyperarticulated context. Figure 4.26 presents box plots of raw lip protrusion values (in mm) for bunched and retroflex tongue shapes according to context (non-hyperarticulated versus hyperarticulated). It suggests that although lip protrusion increases in hyperarticulation across the board, hyperarticulated bunched /r/ is produced with more lip protrusion than hyperarticulated retroflex /r/. The median value of lip protrusion in hyperarticulated retroflex tokens roughly corresponds to that of the non-hyperarticulated

bunched ones. There are however, a larger number of outliers in the retroflex tokens than the bunched ones.

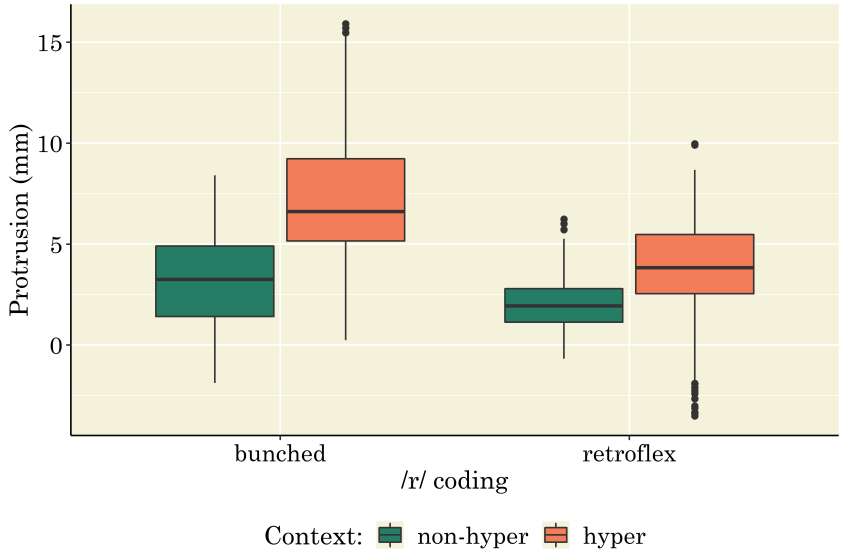



**Figure 4.26:** *Box plots of raw lip protrusion values (in mm) for retroflex and bunched* /r/ *according to context (non-hyperarticulated versus hyperarticulated).*

Figure 4.27 presents mean lip protrusion values for /r/ produced in non-hyperarticulated and hyperarticulated speech in each speaker. The speakers are ordered from most bunched to most retroflex. In the vast majority of speakers, lip protrusion increases on average in the hyperarticulated context. The most substantial increases seem to occur in the first few speakers presented in the graph, i.e., in the speakers who only present bunched tongue shapes. Increased lip protrusion is particularly evident in speakers 17 and 10 both of whom use exclusively bunched tongue shapes.

If hyperarticulation is targeted in order to increase the phonetic distance between the cues distinguishing the target from the competitor, we may observe different degrees of lip protrusion according to the labial features of the competitor. We elicited hyperarticulation by simulating computer recognition errors where word-initial /r/ was recognised as either 'l' or 'w' in the programme's text feedback response. It could be argued that increasing the

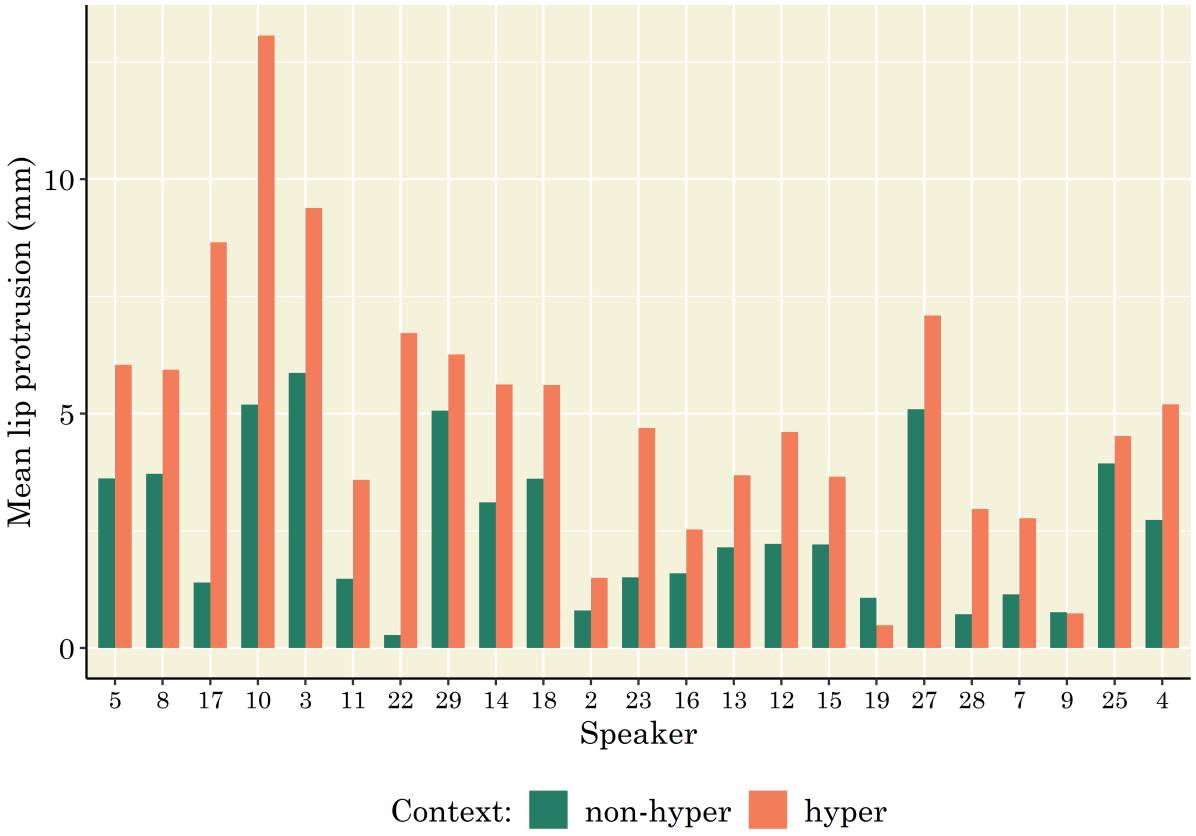



**Figure 4.27:** *Mean lip protrusion (in mm) per speaker according to context (non-hyperarticulated versus hyperarticulated). Speakers are ordered from most bunched to most retroflex.*

degree of lip protrusion for /r/ when placed in direct competition with /w/, which is produced with labialisation, would actually decrease the phonetic distance between the target and the competitor. Whereas, unlike /w/, as /l/ is not produced with labialisation (at least word-initially), lip protrusion for /r/ would increase the /r/-/l/ contrast. Figure 4.28 presents raw lip protrusion values in both non-hyperarticulated and hyperarticulated productions of /r/ according to tongue shape (retroflex or bunched). Hyperarticulated productions were divided into the following three sub-categories:

**Initial hyper**: initial production of target words

/w/ **competitor**: speech repairs directly following a recognition error of 'w'

/l/ **competitor**: speech repairs directly following a recognition error of 'l'

The box plots suggest very little difference in the degree of lip protrusion between /r/ productions correcting 'l' misrecognitions and those correcting 'w' misrecognitions, regardless of tongue shape. We decided not to run statistical analysis comparing the degree of lip protrusion for /r/ between /w/ and /l/ competitors because there was not enough data to do so with any experimental power ($n$=268) and because the box plots show little evidence to suggest a robust difference. The box plots also indicate that the degree of lip protrusion does not greatly differ between the initial productions of target words in the hyperarticulation session and the speech repairs directly following a recognition error, indicating that hyperarticulation was targeted to /r/ productions across the entire session. This is perhaps not surprising given the fact that misrecognitions were never followed by more than four correct recognitions of /r/ across the hyperarticulated session (following the results from Stent et al., 2008). As a consequence, all productions of /r/ in the hyperarticulation session will be pooled in subsequent analyses to increase experimental power.

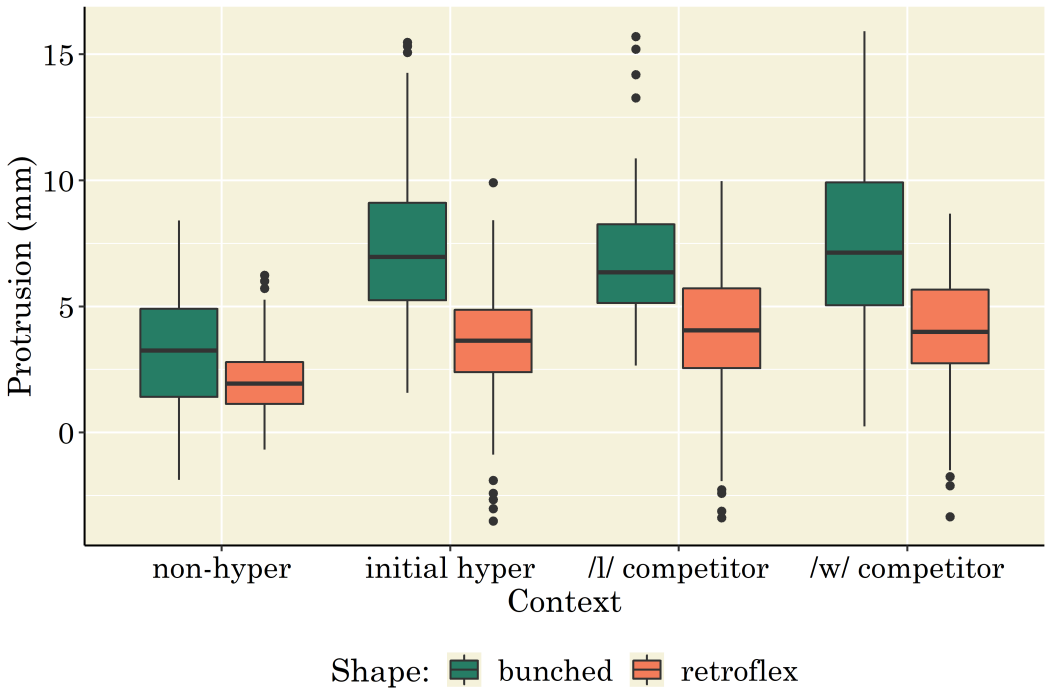



**Figure 4.28:** *Box plots of raw lip protrusion values (in mm) for* /r/ *according to tongue shape and context including competitor information.*

### Acoustics

It was predicted that hyperarticulation would result in lower F3 values than those observed for non-hyperarticulated /r/. The box plots in Figure 4.29 show the effect of hyperarticulation on F3 in bunched and retroflex shapes in female speakers. The median values do not greatly differ between non-hyperarticulated and hyperarticulated contexts in both bunched and retroflex /r/, although they do lower ever so slightly in hyperarticulation.

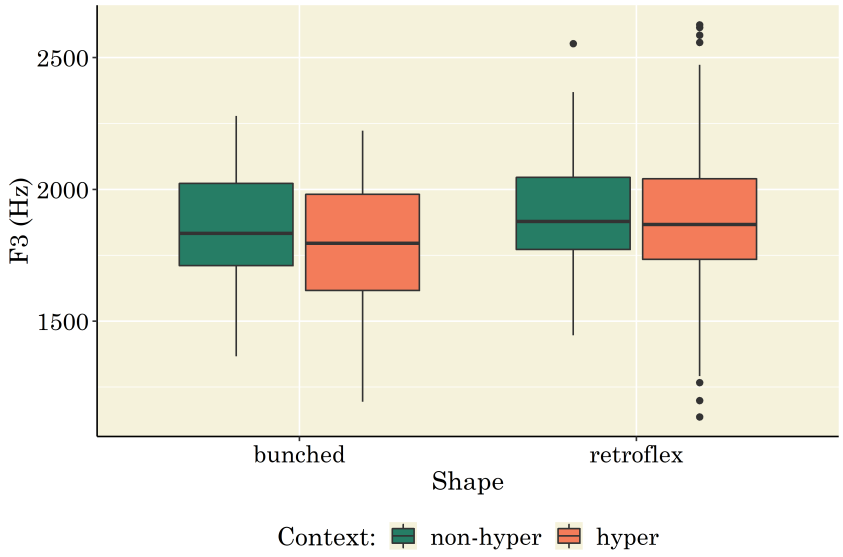



**Figure 4.29:** *Box plots of raw F3 values (in Hz) for bunched and retroflex /r/ in women according to context (non-hyperarticulated versus hyperarticulated).*

Figure 4.30 presents the mean F3 value (in Hz) for all speakers according to context (non-hyperarticulated versus hyperarticulated). Again, speakers have been ordered from most bunched to most retroflex. No obvious trends seem to occur with regards to tongue shape. For the majority of speakers (16/24) F3 decreases on average in the hyperarticulated session. While in some speakers the decrease in F3 is substantial (i.e., F3 drops by over 250 Hz in Speaker 18), decreases to F3 are much subtler in other speakers, and in eight speakers F3 actually increases on average.

Although no predictions were made regarding F2, for the sake of clarity, we present box

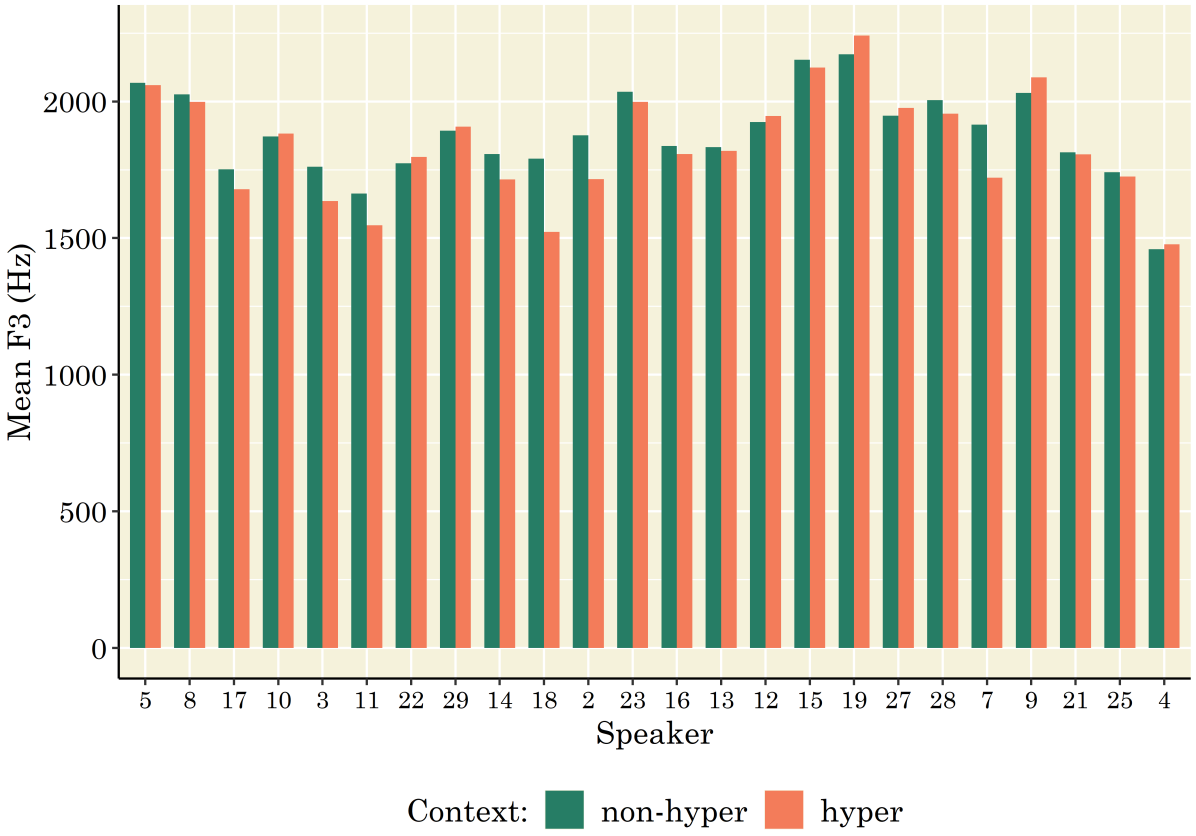



**Figure 4.30:** *Mean F3 (in Hz) per speaker according to context (non-hyperarticulated versus hyperarticulated). Speakers are ordered from most bunched to most retroflex.*

plots for F2 in Figure 4.31. For retroflex tokens, F2 appears to lower in hyperarticulation, while the median F2 value goes up in hyperarticulated bunched tokens. However, both effects appear small.

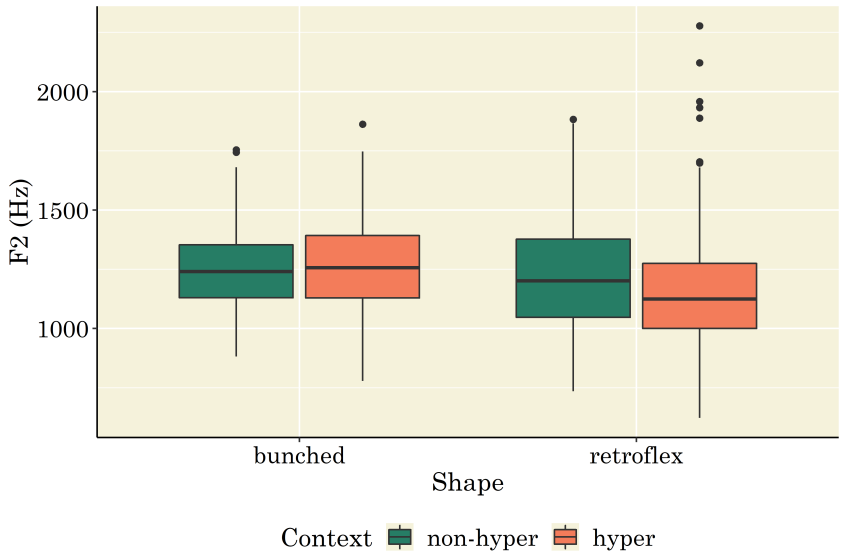



**Figure 4.31:** *Box plots of raw F2 values (in Hz) for bunched and retroflex /r/ in women according to context (non-hyperarticulated versus hyperarticulated).*

### 4.3.5 *Predicting hyperarticulation*

To assess to what extent hyperarticulation may be predicted by tongue shape (retroflex and bunched), lip protrusion and /r/ acoustics (F2 and F3), we performed a generalised linear mixed-effects regression analysis with Context (non-hyperarticulated versus hyperarticulated) as the binary outcome variable. The fixed factors were F3, F2, Protrusion and Shape. An interaction term between Protrusion and Shape was also included. Numeric fixed factors were converted into z-scores by mean centring[9] and then standardising by dividing by the standard deviation. Standardising improves model fit and allows us to measure the relative impact of all variables on the response variable by removing their metric (Winter, 2020). The random structure included

[9]To centre a fixed factor, we subtract the mean of that fixed factor from each data point (Winter, 2020).

by-Speaker and by-Vowel varying intercepts. Likelihood ratio tests revealed that the interaction between Shape and Protrusion was significant ($\chi^2(1) = 10.59$, $p = 0.001$). The main effects of both Shape and Protrusion were also significant (Shape: $\chi^2(1) = 6.16$, $p = 0.01$; Protrusion: $\chi^2(1) = 138.93$, $p < .001$). F3 was also a significant main predictor of hyperarticulation ($\chi^2(1) = 12.78$, $p < .001$). However, F2 did not reach significance ($\chi^2(1) = 1.01$, $p = 0.32$). The final model output presented in Table 4.10 indicates that for an average speaker, the log-odds of an /r/ token being a hyperarticulated one are 0.71 higher when F3 decreases, suggesting that hyperarticulated /r/ has lower F3 values than non-hyperarticulated /r/, following our prediction. Moreover, the log-odds of an /r/ token being hyperarticulated are 1.43 higher when retroflexion increases and 3.97 higher when lip protrusion increases. This suggests that both lip protrusion and retroflexion increase in hyperarticulation but the degree of lip protrusion is particularly impacted. However, the significant interaction between tongue shape and lip protrusion indicates that the model predicts hyperarticulated retroflexes to have less lip protrusion than hyperarticulated bunched /r/ productions.

| Predictor | Estimate (log-odds) | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | -0.52 | 0.76 | -0.69 | 0.50 |
| F2 | -0.15 | 0.15 | -0.15 | 0.32 |
| F3 | -0.71 | 0.20 | -3.48 | $< .001^{***}$ |
| Shape Retroflex | 1.43 | 0.57 | 2.50 | $0.02^{*}$ |
| Protrusion | 3.97 | 0.57 | 6.97 | $< .001^{***}$ |
| Shape Retroflex × Protrusion | -1.82 | 0.60 | -3.00 | $0.003^{**}$ |

$$Context \sim F2\_z + F3\_z + Shape \times Protrusion\_z + (1|Subject) + (1|Vowel)$$

**Table 4.10:** *Output of a generalised mixed effects logistic regression predicting hyperarticulation. Numeric variables were converted to z-scores.*

### 4.3.6 Summary of results

Putting together the various analyses from this chapter, the following findings emerge. Firstly, Anglo-English /r/ may be produced with a range of tongue shapes from curled up retroflex (CU)

to tip down bunched (MB), although retroflexion is more common than bunching. 3 subjects who come from the south east of England produce both retroflex and bunched configurations, while the remaining 21 subjects who come from all over England use either retroflex or bunched shapes. However, given the lack of geographically-stratified data presented here, we cannot comment on any potential regional patterns regarding tongue shape for /r/.

In retroflex users, our results suggest that the degree of retroflexion is related to the quality of the following vowel. The close-front FLEECE vowel appears to be the least compatible with retroflexion, contrary to the open-back LOT vowel. In the three speakers who presented both retroflex and bunched tongue shapes, bunching was only utilised in conjunction with the frontest vowels in the dataset. Although speakers who use exclusively bunched shapes tend to have acquired one distinct tongue shape for /r/, one speaker produces a different, arguably more bunched tongue shape in the context of /r/ followed by the FLEECE vowel. Furthermore, tongue contour tracings revealed that even in speakers who use one distinct shape for /r/, the following vowel has a co-articulatory influence because the tongue is generally more anterior for /r/ followed by the front FLEECE vowel than /r/ followed by the back LOT vowel.

Our analysis suggests that the degree of lip protrusion for /r/ may be related to both tongue shape and the following vowel. According to our statistical analyses, bunched tongue shapes have significantly more lip protrusion in both non-hyperarticulated and hyperarticulated speech. A linear-mixed effects regression model predicted that productions of /r/ followed by the rounded vowels in LOT, THOUGHT and GOOSE have the most lip protrusion of all the vowels in non-hyperarticulated speech, suggesting there is a co-articulatory influence of the labial properties of the following vowel on /r/. However, no significant difference in lip protrusion was observed between /r/ followed by the FLEECE vowel and /r/ followed by the rounded vowels in LOT, THOUGHT and GOOSE, which is unexpected given that the FLEECE vowel is non-rounded.

Finally, hyperarticulated productions of /r/ result in higher degrees of retroflexion and lip protrusion, presumably in order to increase the size of the front cavity, which has a significant lowering effect on F3. F2, on the other hand, is not significantly affected by hyperarticulation.

As in non-hyperarticulated productions of /r/, bunched tongue shapes are accompanied by more lip protrusion than retroflex ones in hyperarticulation.

## 4.4 Discussion

### *4.4.1 Tongue shapes for Anglo-English* /r/

As is the case for the articulation of /r/ in other Englishes, Anglo-English presents a range of possible tongue shapes from tip down bunched (Mid Bunched) to sublaminal (Curled Up) retroflex. However, the production of Anglo-English /r/ differs from the results of recent studies on American English in that retroflexion is much more common in Anglo-English. For example, out of 27 subjects, Mielke et al. (2016) only observed 2 producing exclusively retroflex tokens in both pre- and post-vocalic /r/, compared to our 14/24 subjects in prevocalic /r/. Although their classification would consider our Front Up configuration to be bunched and not retroflex, if we do the same, our Anglo-English data still have far more exclusively retroflex users (25%) than the American English data (<8%). This difference may also reflect the fact that our data are limited to word-initial /r/, whereas Mielke et al. (2016) also included prevocalic /r/ in onset clusters. However, Mielke et al. (2016) observed the highest rates of retroflexion to occur in the same prevocalic syllable-initial context used in the present study. Our results therefore support Hypothesis 1: Anglo-English /r/ is more likely to be produced with retroflexion than American English.

More frequent retroflexion has also been observed in non-rhotic New Zealand English. In a large-scale ultrasound study of 62 New Zealand English speakers, nearly 20% of subjects produced exclusively retroflex tongue shapes (Heyne et al., 2018). Like Mielke et al. (2016), Heyne et al. (2018) considered the equivalent of our Front Up classification to be bunched and not retroflex. If we do the same, the percentage of exclusively retroflex users in Anglo-English (25%) and New Zealand English (nearly 20%) are remarkably consistent. It appears then that exclusively retroflex tongue shapes are up to three times more frequent in non-rhotic than in rhotic Englishes. Heyne et al. (2018) speculated that as New Zealand English speakers very

rarely produce /r/ in post-vocalic environments, where bunching is heavily favoured, speakers are less likely to acquire bunched /r/ as an alternative articulation strategy if they have already mastered retroflexion. Our Anglo-English data seem to support this suggestion. Future studies could consider to what extent the production of /r/ varies in children acquiring rhotic and non-rhotic Englishes.

Although retroflexion is generally more frequent in non-rhotic than in rhotic English speakers, rate of retroflexion is influenced by coarticulation with neighbouring segments. In the present study, our results indicate that retroflexion is favoured by open-back vowels versus close-front ones, in a similar fashion to American English (Mielke et al., 2016; Ong & Stone, 1998; Tiede et al., 2010). The incompatibility of retroflexion with close-front vowels, notably in the FLEECE, KIT and GOOSE vowels, is manifested through the use of less extreme retroflex variants, i.e., less curling back of the tongue tip, less tongue tip raising, and more bunching. Our data therefore support Hypothesis 2: tongue shapes are affected by coarticulation with the following vowel. The shift from extreme retroflexion towards more bunched configurations in close-front vowel contexts further strengthens the argument that the possible tongue shapes for /r/ are on a continuum rather than the initial suggestion (Uldall, 1958) of dichotomous categories. In the present study, speakers who present both retroflex and bunched shapes produce bunched tokens only in the context of a close-front vowel, particularly with the FLEECE vowel. Our results therefore seem to corroborate Hamann (2003)'s suggestion that the tongue shape for [i], which involves the tip being tucked under the lower front teeth, is inherently incompatible with that of retroflexion. Unlike in retroflexes, the tongue tip remains relatively low in the mouth for bunched /r/, which is perhaps why bunching is more compatible with close-front vowels than retroflexion. In one buncher (speaker 10), /r/ preceding all vowels except for the FLEECE vowel were produced with a Front Bunched configuration. /r/ before FLEECE, however, was produced with a Mid Bunched configuration. We observed from tongue contour tracings that the Mid Bunched configuration generally has a lower tongue tip than the Front Bunched one in speakers who present both bunched shapes, which would thus explain why the Mid Bunched shape with a lower tongue tip is preferred in the context of the FLEECE

vowel. It therefore seems natural to consider the Mid Bunched category as the most bunched tongue configuration, despite the fact that bunching, which is generally associated with a dip in the tongue surface, is less apparent than in the Front Bunched shape. We therefore conclude that our continuum ranges from tip down Mid Bunched, most compatible with close-front vowels, to tip up Curled Up retroflex, most compatible with open-back ones.

### 4.4.2 *The contribution of the lips to the production of* /r/

An important finding from the present study is the fact that the degree of accompanying lip protrusion may be influenced by the configuration of the tongue. Bunched tongue shapes have significantly more lip protrusion than retroflex ones. As discussed in Section 2.4, retroflex articulations, by definition, include the addition of a sublingual space, which increases the volume of the front cavity, thus lowering the third formant. Bunched /r/ is produced with the tongue tip positioned relatively low in the mouth and therefore presumably creates less space underneath the tongue tip than retroflex /r/. The difference we observe regarding the degree of lip protrusion could thus be a compensation strategy used by bunchers to lengthen the front cavity in order to obtain the same sized front cavity and therefore, the same acoustic output as retroflexers. These results are thus in line with Hypothesis 5. Indeed, like previous studies on English /r/, we observed no statistically significant difference across tongue configurations in formant values, allowing us to reject the null hypothesis for Hypothesis 3. Unfortunately, formants above F3 were too weak to be tracked in this study. Future studies on Anglo-English could thus attempt to replicate existing studies on other varieties of English such as American English and Scottish English, which have indicated that acoustic differences between tongue shapes exist in the higher formants.

Our analysis also indicates that the use of lip protrusion as a compensation strategy may go beyond the bunched-retroflex distinction. Although our results generally support Gimson (1980)'s observation that /r/ productions in the context of rounded vowels present more lip protrusion than in the context of non-rounded vowels, labial coarticulation cannot account for the fact that in the context of the close-front FLEECE vowel, /r/ has significantly more lip

protrusion than in the context of the more open non-rounded vowels such as those in KIT and DRESS. Labial coarticulation does also not account for the lack of a statistically significant difference in lip protrusion between /r/ followed by the FLEECE vowel and the rounded vowels in LOT, THOUGHT and GOOSE. Visualising tongue contour tracings revealed that /r/ preceding the FLEECE vowel is generally produced with a more anterior tongue position than /r/ preceding LOT, no doubt due to lingual coarticulation. As this fronting of the tongue will presumably result in the shortening of the front cavity, speakers may again compensate for this shortening by increasing lip protrusion, thus extending the front cavity, regardless of underlying tongue shape. A limitation to our analysis is that in the present dataset, place of articulation and rounding are partly confounded: the only non-rounded back vowel is the STRUT vowel, which may actually be realised as the rounded [ʊ] in speakers who do not present the FOOT-STRUT split, i.e., in linguistic Northerners, who as it happens, make up the majority of the dataset ($n$=16). Despite our reservations, compensation strategies for coarticulation with front vowels in retroflexes have been observed in other languages. For example, the vowel /i/ was rounded preceding retroflexes in Wembawemba, an extinct Indigenous Australian language, but not in other vowel contexts (Flemming, 2013). It is interesting to note that despite the higher degree of lip protrusion, /r/ preceding the FLEECE vowel still results in significantly higher F3 values than /r/ preceding all other vowels in the dataset. It seems then that increased lip protrusion does not necessarily result in complete compensation for lingual coarticulation with the FLEECE vowel.

Given the significant differences in lip protrusion we have observed between retroflex and bunched tongue configurations, future studies could consider whether this difference is perceptually salient to an interlocutor in both the auditory and visual domains. Furthermore, although some clues may lie in higher formant values, without the use of advanced and rather expensive instrumental techniques capable of imaging or tracking the tongue, researchers are not yet capable of telling a bunched /r/ from a retroflex one. Visualising the lips, however, can be accomplished with ease, and could therefore be an alternative, more cost-effective strategy. However, we stress that although our data point towards a possible articulatory

compensation strategy involving the use of lip protrusion to extend the front cavity for /r/, more articulatory data, ideally from a more robust imaging technique which would provide vocal tract dimensions i.e., real-time MRI, is evidently required. Indeed, another limitation to our study is the fact that the sublingual space is not visible from ultrasound data. Furthermore, there may well be a three-way trading relation between the size of the sublingual space, palatal constriction location and degree of lip protrusion, which falls outside the scope of this study. Although we have focused on Anglo-English, we see no reason why the use of lip protrusion as a compensation strategy for /r/ could not be extended to other varieties of English, which could also be the object of further study.

Results comparing non-hyperarticulated productions of /r/ to hyperarticulated ones also indicate that lip protrusion is employed to enhance the discriminability of /r/. F3 is a significant predictor of hyperarticulation: the lower the F3 the more likely a production be a hyperarticulated one. However, hyperarticulated bunched tokens are still accompanied by higher degrees of lip protrusion than retroflex ones, which again points towards a relationship between tongue shape and lip protrusion. An alternative hyperarticulation strategy appears to be the use of more extreme retroflex shapes in retroflex users, which would result in an increase in the size of the sublingual space and thus lower F3 values. Indeed, two speakers who used both retroflex and bunched /r/ modified their habitual bunched tongue shape in the context of the FLEECE vowel for retroflex ones in some hyperarticulated tokens. Increased retroflexion is not available to speakers who exclusively bunch their tongue, which may explain why bunchers produce more lip protrusion than retroflexers in hyperarticulated /r/ productions.

Interestingly, although the lowering of all formants, particularly F2, would be the expected acoustic consequence of greater lip rounding, hyperarticulation does not induce significantly lower F2 values for /r/. It seems then that speakers are able to retain the small space between F3 and F2 in their labial articulation of /r/ even when protrusion increases. Although our analysis does not tell us to what extent F3 lowering in hyperarticulation is the result of changes in lip protrusion or in tongue shape, our results suggest that speakers can actively control the articulatory parameters available to them in order to enhance the discriminability of /r/.

Throughout this study, we have assumed that the goal of hyperarticulation is an acoustic one. However, in our 'silent speech' paradigm, speakers may be enhancing intelligibility in the visual domain rather than the auditory one. Indeed, Garnier et al. (2018) considered whether hyperarticulation involves the active enhancement of visible speech cues in Lombard Speech. They found that some speakers (4/6) use visual enhancement when interacting face-to-face with the experimenter in noisy conditions. These strategies were absent when the experimenter turned her back to the speaker. In our study, /r/ was hyperarticulated in the context of /w/ and /l/ competitors. It is perhaps surprising that the articulation of /r/ is enhanced with the lips when placed in direct competition with another sound that has a known labial component, [w]. Indeed, no obvious differences were observed between the degree of lip protrusion in /r/ productions following /w/ competitors and those following /l/ competitors, although we were unable to test the statistical significance of this observation given the limited dataset. If the goal of hyperarticulation is to increase the phonetic distance between the target and its competitor, whether that be in the visual or acoustic domain, the increased labiality we observed in hyperarticulated /r/ may suggest that /r/ has a labial component that contrasts with that of /w/. As a result, the labial gestures for /r/ and /w/ will be directly compared in the next part of this thesis.

## 4.5 CHAPTER CONCLUSION

Articulatory data presented in this study have shown that Anglo-English /r/ is not only produced with a tip up tongue configuration but presents similar lingual variation to that observed in rhotic Englishes with tongue shapes ranging from tip down bunched to curled up retroflex. However, retroflexion is three times more frequent in Anglo-English than in American English, which may be a consequence of the absence of post-vocalic /r/ productions in Anglo-English, a context which favours bunching, as discussed by Heyne et al. (2018). Although some speakers present one configuration exclusively, in others, tongue shape may be directly related to the following vowel with tip up variants favouring open-back vowel contexts and tip down ones favouring close-front ones. A novel finding of this study is that

the degree of accompanying lip protrusion may be directly related to the size of the front cavity in Anglo-English with smaller front cavities presenting the most lip protrusion. Tip down tongue shapes, which have less space underneath the tongue than tip up ones, appear to compensate for their smaller cavity volume through increased lip protrusion. Lingual coarticulation with neighbouring front vowels may reduce the size of the front cavity for /r/ regardless of tongue shape, for which speakers also seem to compensate via increased lip protrusion. When speakers are forced to hyperarticulate their production of /r/, one strategy includes increased lip protrusion, particularly in bunchers. Targeted hyperarticulation of /r/ results in the lowering of F3, which is considered to be the most salient acoustic feature of /r/. We therefore conclude that lip protrusion is an articulatory mechanism used to enhance the saliency of /r/.

# Labialisation in Anglo-English /r/ and /w/

# 5

## 5.1 Introduction

It is well-documented that labiodental productions of /r/ are a common feature of Anglo-English. These variants presumably lack a lingual constriction resulting in higher third formant frequencies than their post-alveolar counterparts (Foulkes & Docherty, 2000). It has been suggested that labiodental variants have emerged by speakers retaining the labial gesture of /r/ at the expense of the lingual one (Docherty & Foulkes, 2001; Foulkes & Docherty, 2000; Jones, 1972), perhaps due to the heavy visual prominence of the lips (Docherty & Foulkes, 2001). These claims suggest two things: firstly, that the labial component of /r/ in this variety is always labiodental even in productions which still have an accompanying lingual gesture; and secondly, that the labial gesture is visually prominent. As the exact phonetic implementation of labialisation in /r/ is unknown, these assumptions have yet to be confirmed and therefore warrant further study. We intend to verify both of these claims in this thesis, starting here with the idea that /r/ productions with an observable tongue gesture are produced with a labiodental lip configuration. A detailed description of the lip posture accompanying lingual productions of approximant /r/ may give us some indication as to why increased labiodentalisation has

occurred. If /r/ is labiodental, the labial gesture for /w/, which is unequivocally considered rounded, should differ considerably. We therefore aim to compare the lip postures for /r/ and /w/ using the video camera data we collected for Experiment 1.

The results from our hyperarticulation study in Experiment 1 gave some evidence to suggest that the lip configuration for /r/ may indeed differ from that of /w/. Hyperarticulation was elicited by getting participants to correct recognition errors where /r/ was mistaken for 'l' or 'w' by a simulated automatic speech recognition programme. If the goal of hyperarticulation is to enhance the phonetic distance between the target and the competitor, the increased labiality we observed in hyperarticulated /r/ seems to suggest that /r/ has a labial component which contrasts with that of /w/. Furthermore, the resulting formant values of hyperarticulated /r/ suggest that /r/ may not necessarily be produced with lip rounding. The only significant acoustic predictor of hyperarticulation was F3, which is generally considered to be the most salient acoustic feature of /r/. If increased lip rounding (i.e., involving a decrease in lip area) is a concomitant of increased lip protrusion, we would expect significant decreases to F2, which was not the case. Increased labiality may thus allow speakers to lower F3 for /r/ while maintaining a small distance between F2 and F3. Indeed, researchers have remarked on the close proximity of F3 to F2 for English /r/ (Dalston, 1975; Guenther et al., 1999; Lisker, 1957; O'Connor et al., 1957; Stevens, 1998). The results from the previous study therefore indicate that /r/ has a labial component which not only enhances the acoustic salience of /r/, but may also contrast with that of /w/. We therefore derive the following hypothesis:

**Hypothesis 6** /r/ has a specific lip posture which differs from that of /w/ in Anglo-English.

### *5.1.1 Principal phonetic properties of /r/ and /w/*

A detailed phonetic description of the English approximant /r/ has been supplied in Chapter 2. To summarise, post-alveolar articulations of /r/ are typically produced with three simultaneous constrictions in the vocal tract: in the pharynx, in the mid-palatal region and at the lips. The lingual constriction may be produced with a multitude of tongue shapes from curled-up retroflex to tip down bunched. A common characteristic of the various possible shapes is a

large front cavity (between the lips and the palatal constriction). This large front cavity has been associated with /r/'s most salient acoustic property: a very low third formant in close proximity to F2. Like /r/, /w/ is also a sonorant consonant in English meaning that it too has a characteristic formant pattern. It shares similar acoustic features to its vocalic counterpart /u/, notably with a very low F2. Like /r/, /w/ is produced with multiple constrictions: one at the lips and one at the palate in the velar region. The attainment of its particularly low F2 is associated both with its labial and with its palatal constriction. As Vaissière (2009) remarked, when the constriction is in the back section of the vocal tract, 'rounding/protrusion and backing of the tongue form a single functional entity, which has a single acoustic correlate: a low F2' (p. 29). The removal of /w/'s labial component would result in an increase in F2. In his acoustic modelling of vowels, Stevens (1998) explained that in the case of a backed tongue constriction, the condition of minimum F2 is achieved 'only if there is a constriction at the lips, that is, if the lips are rounded and a narrow opening is formed' (pp. 281-282). Similarly, the example of Fant's nomograms for a narrow lingual constriction presented in Figure 3.3 (Chapter 3, p. 86) indicates that the lowest possible F2 values coincide with the smallest degree of lip opening, regardless of the front-back position of the tongue. Acoustic modelling therefore suggests that close lip rounding, or horizontal labialisation, is a requirement for /w/ in order to attain its characteristically low F2. For /r/, Experiment 1 indicated that lip protrusion increases the size of the front cavity, which contributes to the lowering of F3. However, the width and height of the lips for /r/ have yet to be considered. Finally, the acoustic profile of /w/ and post-alveolar /r/ should differ significantly: /r/ should have a significantly lower F3 but a significantly higher F2 than /w/.

As far as we are aware, very few articulatory accounts of labiodental approximants exist. This is probably because they are particularly rare in the world's languages. As Gick et al. (2019) noted, a labiodental approximant is reported to occur in only 6 of the 451 languages in the UCLA Phonological Segment Inventory Database (UPSID) (Maddieson, 1984; Maddieson & Precoda, 1989). Ordinarily in labiodentals, the lower lip moves towards the top teeth (Ladefoged & Maddieson, 1996). However, the area where the narrowest constriction occurs between the

bottom lip and top teeth may vary. Although one might presume that for a typical labiodental fricative such as [f] and [v], the bottom lip is retracted to bring the lower lip back over the lower teeth and the bottom lip comes into contact with the backs of the upper incisors, Ladefoged and Maddieson (1996) observed that speakers of English tend not to present a large degree of bottom lip retraction. Instead, the bottom lip is positioned so that the narrowest constriction occurs between the inner surface of the bottom lip and the front surface of the incisors. These two labiodental articulations were described by Catford (1977) as exolabial and endolabial respectively, with either the outer or inner surface of the lips forming the place of articulation. He presents profile view schematisations of labiodental articulations, which have been recreated in Figure 5.1. A schematisation of an 'endolabial-endolabial' articulation, which Catford (1977) associates with the 'rounding' employed for [w], has also been included for ease of comparison between the three labial articulations. The schematisations indicate that an exolabial-dental articulation, where the bottom lip is retracted, would essentially decrease the size of the vocal tract. Given what we know about the use of lip protrusion for /r/ from Experiment 1 and its role in extending the size of the cavity in front of the lingual constriction, this exolabial-dental articulation seems entirely incompatible with /r/. However, the endolabial-dental articulation remains a possibility. As the narrowest constriction between the upper incisors and the bottom lip occur *inside* the mouth, the bottom lip is free to extend outward, thus increasing the size of the front cavity. Although these schematisations are of course simplistic, we notice that the front-back position of the bottom lip for the endolabial-dental articulation is not dissimilar to that of the endolabial-endolabial one, associated with [w].

As far as we are aware, no empirical study currently exists which specifically measures the lip posture of labiodental articulations. As a result, based on previous studies which present measures of lip rounding (principally in vowels), we intend to compare the lip postures for /r/ and /w/ based on three main dimensions: lip protrusion, lip width and lip height. We predict that /r/ is inherently more labiodental than /w/ and as such, their lip postures should differ.


**Figure 5.1:** *Schematised profile views of two labiodental articulations and rounding for* [w] *(adapted from Catford, 1977, Figure 39).*

## 5.2 Methodology

### *5.2.1 Stimuli*

The non-hyperarticulated productions of /r/ and /w/ from Experiment 1 were considered in this study. Speaker 07 was excluded because the camera angle in the frontal lip recordings made it difficult to view her top lip. We therefore present data from 23 speakers who produced 18 /r/-/w/ minimal pairs in isolation. The stimuli are presented in Table 5.1. For a presentation of the experiment procedure and participants, see the methodology of Experiment 1 presented in Chapter 4 (Section 4.2, p. 105).

| Lexical set | /r/-initial | /w/-initial |
|---|---|---|
| FLEECE | *reed* | *weed* |
| FLEECE | *reap* | *weep* |
| DRESS | *red* | *wed* |
| GOOSE | *room* | *womb* |
| KIT | *ring* | *wing* |
| TRAP | *rack* | *whack* |
| STRUT | *run* | *one* |
| THOUGHT | *raw* | *war* |
| LOT | *rot* | *what* |

**Table 5.1:** *Test words including minimal pairs contrasting* /r/ *and* /w/ *word initially and their corresponding lexical set vowel.*

### 5.2.2 Acoustic analysis

The acoustic data were exported as wav files from AAA and analysed in Praat (Boersma & Weenink, 2019). /rV/ and /wV/ (where V corresponds to one of the eight lexical set vowels) were manually annotated as a whole. Praat's Burg algorithm was used to obtain formant values. For each word, formant parameters were manually adjusted in order to reach an optimal match between formant estimation and the underlying spectrogram using the same technique as the one described for Experiment 1 (presented in Chapter 4, Section 4.2.5, p. 115). We chose to consider the most salient acoustic feature of /r/ and /w/ for formant extraction, i.e., F3 and F2, respectively. The minimum F3 value for /r/ and the minimum F2 value for /w/ were extracted by selecting the portion of the recording corresponding to /r/ and /w/ and opening the formant listing in Praat. The point at which F3 was minimally low for /r/ and F2 was minimally low for /w/ were labelled, and the first three formants (F1-F3) were extracted at these points. Thirteen tokens were not analysed due to formant visualisation issues, predominantly due to the use of creaky voice, yielding formant values from a total of 401 tokens.

### 5.2.3 *Measuring the lips by hand*

Lip protrusion was measured from profile lip camera videos of /r/ and /w/ in AAA. One profile lip image was selected for each token corresponding to maximal lip protrusion by holistically examining sequential frames. One image corresponding to a neutral lip configuration (with the lips closed) prior to speech was also selected per speaker. Lip protrusion for /r/ and /w/ was measured using the same technique as in Experiment 1 (presented in Chapter 4, Section 4.2.7, p. 124), i.e., by calculating the distance from a neutral lip position to maximum lip protrusion. As the front and profile lip cameras were synchronised, the corresponding front view images were used to measure lip height and lip width.

Lip height and width for /r/ and /w/ were measured at the same time point as the lip protrusion measure, i.e., at the point of maximum lip protrusion. Width and height were considered because our review of phonetic studies presented in Chapter 3 indicated that labialisation is predominantly implemented via changes to the horizontal (i.e., width) and vertical (i.e., height) lip setting. Lip measurements were inspired by those presented in Garnier, Ménard, and Richard (2012) and Mayr (2010), where lip width was measured at the lip corners, and height was measured from the middle of the top lip to the middle of the bottom lip.

For lip width, a fiducial line was positioned to coincide with the quasi-horizontal line which is naturally formed between the top and bottom lip when the lips are closed in a neutral position. This horizontal fiducial ran parallel to the upper and lower edges of the video pane. A vertical line was then positioned at each lip corner intersecting the horizontal fiducial, as presented in Figure 5.2. Using AAA, we calculated the distance between the left and right lip corner along the horizontal lip fiducial in the neutral front image and in /r/ and /w/. To quantify lip height, another lip fiducial was positioned to vertically dissect the lips approximately at their mid-point at the philtrum dimple in their neutral setting. This fiducial line ran parallel to the left and right edges of the video pane. A horizontal line was positioned at the vermilion border of the outer edge of the top and bottom lip intersecting the vertical fiducial, as presented in Figure 5.2. Using AAA, we calculated the distance between the top and bottom lip along the vertical lip fiducial in the neutral front image and in /r/ and /w/. Each speaker was assigned

one horizontal fiducial and one vertical fiducial, which were used for all his/her lip height and lip width measures. Unlike for the lip protrusion measure (as presented in Chapter 4, Section 4.2.7, p. 124), no scaling device was used for the frontal lip view measurements. The measurements are therefore not in world units. As a result, all manual lip measurements (i.e., lip protrusion, lip height and lip width) were transformed into the percentage of change relative to each speaker's neutral lip setting dimensions, using the following calculation:

$$\text{percentage change} = \left(\frac{\text{maximum lip} - \text{neutral lip}}{\text{neutral lip}}\right) \times 100$$

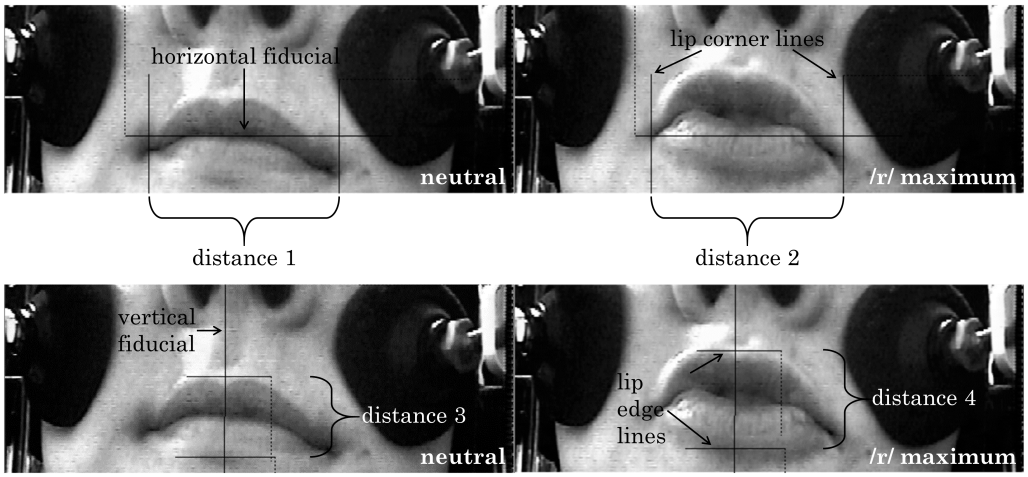



**Figure 5.2:** *Front view manual lip measures. Lip width (distance 2) and lip height (distance 4) were calculated as the percentage change from the neutral lip setting (distance 1 and distance 3, respectively).*

### 5.2.4 Measuring the lips automatically

Measuring lip dimensions by hand is particularly time consuming and may be prone to human error. To make matters worse, video acquisition was done in somewhat adverse conditions.

Camera angle could not be controlled[1] and lighting conditions were not optimised. Measuring the lips from the front camera images was particularly troublesome and as we have previously mentioned, one speaker had to be excluded because her top lip was not visible. Two example images are provided in Figure 5.3. The top image is of Speaker 07 whose data was excluded and the bottom image shows one of the higher quality images in the dataset from another speaker. Given the limitations of hand measures, automatic extraction of the lip contour would evidently be a more reliable, reproducible and less time-consuming approach. As a result, efforts were made to find a technique capable of segmenting the lips from the rest of the image automatically. Attempts to use colour segmentation proved unsuccessful due to poor image quality. It was therefore suggested that we explore the possibility of utilising techniques from deep learning. Indeed, Deep Neural Networks (DNNs) have enjoyed great success in recent years in the field of automatic image recognition (Simonyan & Zisserman, 2015). Although phonetic studies employing such techniques are rare (Ferragne, Gendrot, & Pellegrini, 2019; King & Ferragne, 2019), given the visual nature of the dataset, it seemed like a good opportunity to apply deep learning-based methods to answer phonetic questions.

The most common class of DNNs applied to image classification and recognition is Convolutional Neural Networks (CNNs). The technical details concerning the inner workings of CNN architectures go far beyond the scope of this thesis. However, for image recognition, the idea behind them is relatively straightforward. To put it simply, the aim is to replicate the basic human skill of recognising and classifying objects within an image. For example, if a person were presented with an image of a cat in a field, their brain would automatically recognise the object and classify it as 'cat'. The brain is also able to distinguish a cat from another object or animal, such as a dog. By providing the computer programme with lots of images of cats and dogs in a variety of different settings, the programme should be able to learn the qualities that distinguish the two animals in an image. So, when the computer is presented with a new image of a cat, it should be able to tell with a certain degree of certainty that there is a cat, and

[1]The ultrasound stabilisation headset (Articulate Instruments Ltd., 2008) was not originally conceived to include front and profile lip camera brackets. As a result, although both cameras are stabilised in relation to the speaker's head, adjustments to the angle of the camera are limited.

**Figure 5.3:** *Examples of front camera images of varying quality. The top image comes from Speaker 07 whose data was excluded from this analysis given the poor positioning of the lip camera.*

not a dog, in the image. The 'convolutions' filter the images pixel by pixel. The pixels that are important for a cat to be classified as a cat will be 'enhanced' by the model, whereas non relevant pixels for the cat class will receive negligible weight.

If we return to our initial goal of finding a technique capable of automatically segmenting the lip contour from the rest of the image, we can now imagine a situation whereby a CNN might be capable of learning the features that distinguish the lips from the rest of the face. A technique called *semantic segmentation* was applied to teach a CNN to detect the lip area using Matlab Computer Vision Toolbox (Mathworks, 2020a) and Deep Learning Toolbox (Mathworks, 2019). The front view image corresponding to maximum lip protrusion was manually located and extracted from the 414 lip videos in our dataset, resulting in 207 8-bit colour images of /r/ and of /w/ of size 300 pixels (height) × 800 pixels (width). The 414 lip images were the same as the ones used for the hand measures of lip height and width. 100 of the 414 images were randomly selected and the lip area was manually segmented. Manual segmentation involves labelling the pixels within an image which correspond to a particular object or class. In our

data, we had two classes: the mouth and everything else (the background). A DeepLab v3+ (L.-C. Chen, Zhu, Papandreou, Schroff, & Adam, 2018) CNN based on ResNet-18 (He, Zhang, Ren, & Sun, 2016), a well-known CNN architecture in image recognition, was trained using 60 of the 100 segmented images with their corresponding pixel labels. The remaining 40 images were used to test the model on what it had learnt. For each image, the CNN selects the pixels it has learnt to associate with the lip area, which are then compared with the pixel values obtained from manual segmentation. The model's performance can thus be evaluated. We used the following metrics to evaluate model performance: global accuracy, mean accuracy, mean intersection over union (IoU), weighted IoU and mean boundary F1 (BF) score. Table 5.2 gives a brief description of each metric based on the descriptions presented in Mathworks (2020a) and Costa, Campos, de Aquino e Aquino, de Albuquerque Pereira, and Costa Filho (2019).

Model performance was evaluated at a global and at a class level. Global evaluation metrics are presented in Table 5.3 and Table 5.4 presents metrics for the mouth and the background classes separately (i.e., at the class level). Global accuracy was very high (94.29%) suggesting that the CNN performed very well. As mean IoU penalises false positive predictions, it can be considered to be a more precise measure of performance. Although mean IoU is lower than global accuracy, a mean IoU of 80.79% still suggests that the model performed well. However, the global mean BF score is comparatively lower at 56.23% indicating that the model performed less well at detecting the boundary between the two classes. Indeed, if we consider the class metrics in Table 5.4, the mean BF score for the mouth is only 29.84%. These results suggest that globally, the model was able to segment the mouth from the background but was less successful at detecting the boundary between them i.e., the lip contour.

Given the high global accuracy achieved by the CNN, the resulting model was used to automatically detect the mouth in all 414 front view images. An example image of the resulting automatic segmentation is presented in Figure 5.4. The mouth is presented in blue. Despite the high accuracy score, automatic segmentation of the mouth does present stray pixels, although the CNN was generally able to localise the mouth quite well. In order to prevent bias caused by

| Metric | Description |
|---|---|
| Global accuracy | Ratio of correctly classified pixels to the total number of pixels. |
| Mean accuracy | Ratio of correctly classified pixels in each class to the total number of pixels, averaged over all classes. |
| Mean IoU | Ratio of correctly classified pixels to the total number of pixels that are assigned that class by manual segmentation and by the predicted one, averaged over all classes. Penalises the incorrect classification of pixels as the mouth (false positive) or as background (false negative). |
| Weighted IoU | Average IoU of all classes in the image, weighted by the number of pixels in each class. Used when there is a disproportionate relation between the class sizes in the images, minimising the penalty of wrong classifications in smaller classes. |
| Mean BF score | Measures how close the predicted boundary of an object matches the manually segmented boundary. Mean BF score measures the average BF score of all images. |

**Table 5.2:** *Evaluation metrics for semantic segmentation using a CNN.*

| Global accuracy | Mean accuracy | Mean IoU | Weighted IoU | Mean BF Score |
|---|---|---|---|---|
| 0.9429 | 0.9518 | 0.8079 | 0.9029 | 0.5623 |

**Table 5.3:** *Global evaluation metrics for semantic segmentation of the mouth from front camera images using a CNN.*

| Class | Accuracy | IoU | Mean BF Score |
|---|---|---|---|
| mouth | 0.9637 | 0.6808 | 0.2984 |
| background | 0.9399 | 0.9350 | 0.8263 |

**Table 5.4:** *Class evaluation metrics for semantic segmentation of the mouth from front camera images using a CNN.*

these stray pixels, an ellipse was fitted to the region identified as the mouth in each image[2], an example of which is presented in Figure 5.5. The ellipse then allowed us to compute four measurements of the lips (in pixels). These measures were based on the length of the horizontal and vertical axes and the position of the ellipse centroid (i.e., where both axes meet). The

measures and their corresponding lip dimensions are presented in Table 5.5. The first two dimensions (mouth width and height) are comparable to the front view measures we took by hand, as presented in Section 5.2.3. The latter two concern the position of the mouth and are new additions, which emerged somewhat fortuitously from the inclusion of an ellipse.

| Dimension | Ellipse measure (in pixels) |
|---|---|
| mouth width | length of the horizontal axis |
| mouth height | length of the vertical axis |
| horizontal mouth position | position along x-axis of centroid |
| vertical mouth position | position along y-axis of centroid |

**Table 5.5:** *Ellipse measures and their corresponding lip dimensions resulting from automatic semantic segmentation of the lips using a CNN.*

### 5.2.5 Statistical analysis

Linear mixed-effects models were implemented in R (R Core Team, 2018) using the same technique as the one detailed for Experiment 1 (presented in Chapter 4, Section 4.2.8, p. 125) to compare the lip dimensions and the acoustics of /r/ and /w/.

We supplemented the regression analysis of lip measurements with a further CNN analysis, which was employed to automatically classify /r/ and /w/ tokens from our 414 front camera lip images. In a way, we can consider the results from this CNN analysis as an alternative to inferential statistics (Ferragne, 2019). If the CNN is able to classify /r/ and /w/ with a high level of accuracy, we may conclude that /r/ and /w/ present sufficiently discriminant features which allow the programme to distinguish between them.

Unlike the technique we employed to segment the lips from the rest of the image (presented in Section 5.2.4), prior segmentation of the lips was not required for the classification of /r/ verus /w/. The well-known CNN architecture ResNet-18 (He et al., 2016) was used with the input resized to match the size of the images in our dataset (300 × 800 pixels), using Matlab

[2]The 'regionprops' function in Matlab Image Processing Toolbox (Mathworks, 2020b) was used to compute ellipse parameters.

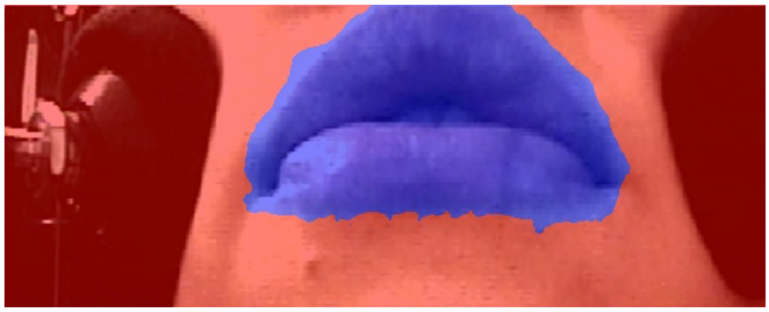


**Figure 5.4:** *Automatic segmentation of the mouth (in blue) via semantic segmentation using a CNN.*

**Figure 5.5:** *Ellipse fitted to the automatically segmented mouth, which is used to compute mouth width, height, and centroid.*

Deep Learning Toolbox (Mathworks, 2019). Two types of model validation were applied: 10-fold cross validation and a type of leave-one-out validation. For 10-fold cross validation, the dataset is randomly split into 10 equal-sized subsets. One subset is put aside for the test stage while the remaining nine are used to train the model. Cross-validation is repeated 10 times with each of the 10 subsets used once for testing. The average classification score is computed across the 10 repetitions to produce one single model estimation. Given that the dataset is split randomly, data from all speakers is present in both training and testing stages of the model, which means that the model may rely on speaker-specific information to make its decisions.

Although this technique is methodologically valid, in order to challenge the generalisation ability of our model, a leave-one-out validation procedure was also employed, whereby, at each step, a speaker's whole dataset is left out for testing, and training is carried out with the data from the remaining 22 participants.

It is often remarked that one of the shortcomings of deep learning is the difficulty in understanding what exactly DNNs learn from the data. Our model may be able to recognise a /r/ from a /w/ with high accuracy, but how do we know that it based its decisions on linguistically relevant information, i.e., the configuration of the lips? Indeed, as Ferragne et al. (2019) pointed out, DNNs are often described as 'black boxes' due to the opaqueness of their inner mechanisms. Luckily for us, solving this problem has been the focus of many researchers in the deep learning community and as a result, effective methods of visualising what DNNs learn now exist (Ferragne, 2019). One such technique is occlusion sensitivity (Zeiler & Fergus, 2014), whereby a mask is placed to cover a small area of each image and the resulting drop in the probability that the image will be correctly classified is recorded. The mask position is then changed slightly and the probability drop of the new mask position is computed until the mask has occluded all possible positions in the image. Matlab's defaults were used for occlusion analysis. Mask size was 60 pixels (height) × 160 pixels (width) and step size (aka 'stride') was 30 × 80. The resulting occlusion analysis may be visualised by overlaying each image with a heatmap showing the areas on which the models based their decisions. It is hoped that the resulting occlusion analysis will reveal that the models rely on the lip area to classify /r/ and /w/, which will be evident from the resulting heatmaps.

## 5.3 Results

### *5.3.1 Acoustics of* /w/ *and* /r/

As our dataset contains limited data from male subjects (n=2) and it is well established that speaker sex influences formant values, we will only consider data from the remaining female subjects (n=21) in our acoustic analysis. Table 5.6 presents average first, second and third

formant values in women for /r/ and /w/ along with their standard deviations. As expected, /r/ has a much lower F3 on average than /w/ (around 800 Hz lower), while /w/ has a much lower F2 than /r/ (around 500 Hz lower). The frequency of F1 is extremely similar between the two sounds. Figure 5.6 presents box plots of F2 and F3 for /r/ and /w/, which paint the same picture. For /w/, we observe a notable distance between F2 and F3 with a very large space between their median values. For /r/ the distance between F2 and F3 is smaller than for /w/. However, the interquartile range (presented by the boxes) of F2 and F3 for /r/ do not overlap, which suggests that F3 is still quite distinct from F2.

| Phoneme | F1 | F2 | F3 |
|---|---|---|---|
| /w/ | 401 (66) | 743 (171) | 2716 (241) |
| /r/ | 418 (65) | 1242 (226) | 1900 (212) |

**Table 5.6:** *Mean formant values (in Hz) and their standard deviations (in parentheses) for* /w/ *and* /r/ *in 21 female subjects.*

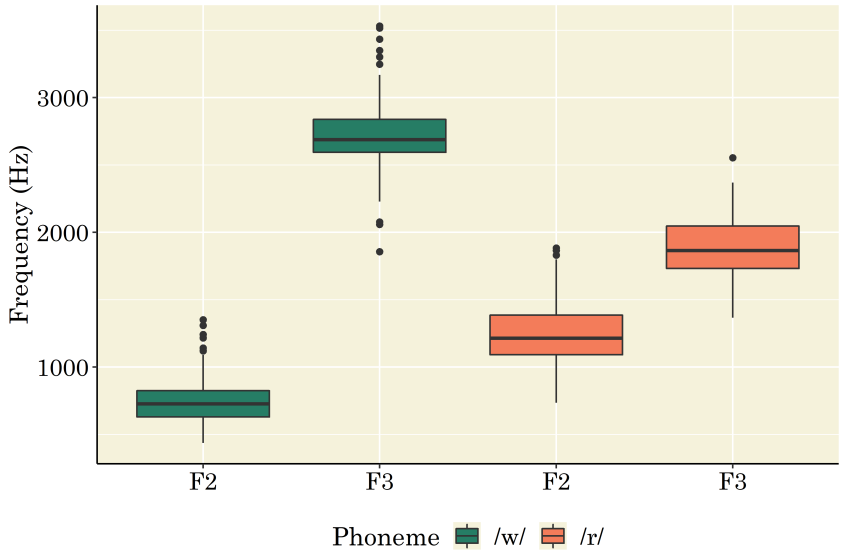



**Figure 5.6:** *Box plots presenting raw F2 and F3 frequencies (in Hz) for* /w/ *and* /r/ *in female subjects.*

We performed a generalised linear mixed-effects analysis in which the first, second and third formants were used to predict the probability that a token is /w/ in the 21 female subjects. Phoneme category (/r/ versus /w/) was thus the binary response variable. The fixed factors (F1, F2, F3) were mean-centred and then standardised (i.e., z-scored) to allow us to measure the relative impact of all formants on the response variable. The random structure included by-Speaker and by-Vowel varying intercepts. Likelihood ratio tests revealed that while F1 did not reach significance ($\chi^2(1) = 0.99, p = 0.32$), both F2 and F3 were significant main effects (F2: $\chi^2(1) = 16.20, p < .001$; F3: $\chi^2(1) = 22.05, p < .001$). The model output presented in Table 5.7 indicates that for an average speaker, the log-odds of a token being a /w/ are 82.50 higher when F2 decreases and 121.44 higher when F3 increases. These results reflect what we observed in the descriptive statistics presented above: /r/ has a lower F3 but a higher F2 than /w/. F2 and F3 are very strong predictors of phoneme category and have similar predicted $t$ values suggesting that both formants are prominent acoustic cues for the /r/-/w/ contrast.

| Predictor | Estimate (log-odds) | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 22.32 | 7.99 | 2.79 | $< .01^{**}$ |
| F1 | -7.89 | 5.28 | -1.50 | 0.14 |
| F2 | -82.50 | 19.74 | -4.18 | $< .001^{***}$ |
| F3 | 121.44 | 29.54 | 4.11 | $< .001^{***}$ |

*Phoneme ~ F1_z + F2_z + F3_z* + (1|*Subject*) + (1|*Item*)

**Table 5.7:** *Output of a generalised linear mixed-effects model predicting the probability a token is a* /w/ *according to the first three formants. Formant values (F1 to F3) were converted to z-scores.*

### 5.3.2 *Labial properties of* /w/ *and* /r/

#### Hand measures

As lip width and height were not measured in world units, all three lip dimensions were transformed into the percentage change relative each speaker's neutral lip setting. Table 5.8 presents mean and standard deviation percentage change values from a neutral setting in lip protrusion, width and height in 23 speakers according to phoneme. On average, /r/ and /w/ involve an increase in lip protrusion and height compared to a neutral lip setting, although lip protrusion and height are greater in /w/ than in /r/. The most striking difference between /r/ and /w/ occurs at the lip width. While lip width for /r/ virtually does not change from the neutral setting (less than 0.1% on average), it decreases by nearly 12% on average for /w/. Figure 5.7, which presents box plots of the same data, again indicates that lip width barely changes from the neutral setting for /r/. The median lip width value for /r/ is around 0% and variability in the data is low given the extremely small interquartile range. The fact that lip width decreases from the neutral setting for /w/ is an indication that labialisation for /w/ is produced by drawing the lip corners together i.e., with horizontal labialisation. As there is little change in lip width between the neutral setting and labialisation for /r/, it seems unlikely that horizontal labialisation takes place.

We observe both from the box plots in Figure 5.7 and the standard deviation values in Table 5.8 that lip height is quite variable in both /w/ and /r/. Lip height was particularly

challenging to measure, as the vermilion border of the top and bottom lip is not always apparent in some speakers, which may explain the high variability observed in this measure in comparison to the other two.

| Phoneme | Protrusion | Width | Height |
|---|---|---|---|
| /w/ | 18.65 (11.91) | -11.89 (8.12) | 24.95 (20.25) |
| /r/ | 13.02 (10.15) | 0.09 (3.76) | 16.27 (14.75) |

**Table 5.8:** *Mean and standard deviation (in parentheses) percentage change from a neutral lip posture in lip protrusion, width and height for* /w/ *and* /r/ *according to manual lip measures.*

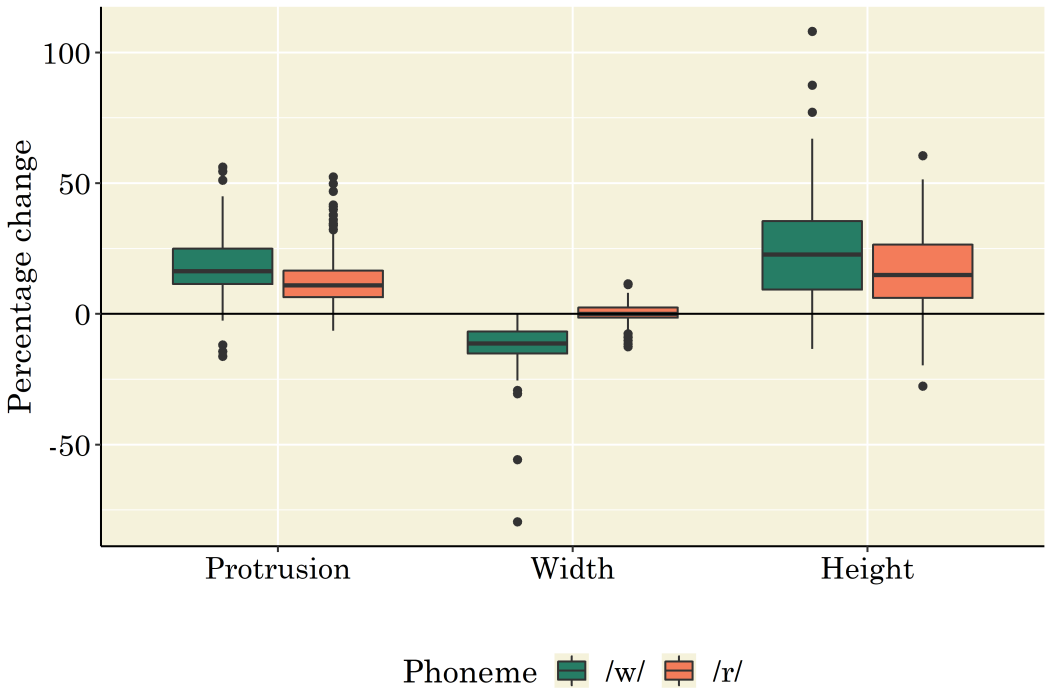



**Figure 5.7:** *Box plots of percentage change from a neutral lip posture in lip protrusion, width and height for* /w/ *and* /r/ *from manual lip measures.*

A generalised linear mixed-effects analysis was performed predicting the probability that a token is a /w/ based on lip protrusion, height and width. In this model, as all three lip measures share the same metric (i.e., percentage change from the neutral setting), we chose

not to centre or standardise them. Raw percentage values were therefore included directly in the model. The random structure had varying intercepts by-Speaker and by-Vowel (with eight levels according to the following lexical set vowel). Likelihood ratio tests revealed that lip width was the only statistically significant main predictor of phoneme ($\chi^2(1) = 452.11, p < .001$). The other lip dimensions did not reach significance (Protrusion: $\chi^2(1) = 1.61, p = 0.2$; Height: $\chi^2(1) = 0.99, p = 0.32$). The model output presented in Table 5.9 indicates that for an average speaker, the log-odds of observing a /w/ token are 3.72 higher when lip width decreases. These results suggest that, based on the three lip dimensions, lip width is the best predictor of phoneme: /w/ has a significantly smaller lip width than /r/, while protrusion and lip height do not significantly differ between /r/ and /w/.

| Predictor | Estimate (log-odds) | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | -19.56 | 9.90 | -1.98 | 0.05 |
| Protrusion | 0.23 | 0.22 | 1.06 | 0.30 |
| Width | -3.72 | 1.56 | -2.38 | 0.02* |
| Height | -0.07 | 0.08 | -0.94 | 0.35 |

*Phoneme ~ Protrusion + Width + Height* + (1|*Speaker*) + (1|*Vowel*)

**Table 5.9:** *Output of a generalised linear mixed-effects model predicting the probability a token is a /w/ according to hand measured lip dimensions.*

Figure 5.8 presents mean lip width values for each speaker for /w/ and /r/ measured as the percentage change from a neutral lip position. On average, lip width for /w/ goes down in comparison to the neutral setting in all 23 speakers, suggesting that a decrease in lip width is a robust characteristic of /w/. Lip width is much more variable for /r/: 13 speakers have wider lips for /r/ than for their neutral position on average, while in other speakers lip width decreases. Regardless of whether or not speakers increase or decrease lip width for /r/, /w/ is always produced with a smaller average lip width than /r/ in all speakers.


**Figure 5.8:** *Mean percentage change from a neutral lip setting in lip width per speaker for* /w/ *and* /r/.

### Automatic measures

Table 5.10 presents descriptive statistics for the four automatic lip measures acquired via semantic segmentation using a CNN. Mean values (in pixels) along with their standard deviations have been included for the four lip measures. The results for lip width and height follow those from our hand measures: /w/ has a smaller lip width than /r/, while lip height does not seem to greatly differ on average between the two. With regards to the horizontal position of the lips, there is unsurprisingly very little difference between the /w/ and /r/. We had no reason to believe that labialisation would result in the lips being positioned more to one side than the other, so this result was expected. However, the mean values for vertical position suggest that /r/ and /w/ vary along this dimension. It is important to specify here that lower values in vertical position are closer to the top of the image. /r/ has a lower vertical lip position than /w/ on average, suggesting that the lips are higher for /r/ than for /w/.

| Phoneme | Width | Height | Horizontal position | Vertical position |
|---|---|---|---|---|
| /w/ | 289.53 (34.47) | 166.50 (23.60) | 375.40 (29.44) | 138.74 (31.97) |
| /r/ | 316.42 (33.29) | 163.96 (26.29) | 374.26 (28.65) | 129.21 (32.95) |

**Table 5.10:** *Mean and standard deviation (in parentheses) lip dimensions (in pixels) for* /w/ *and* /r/ *from automatic semantic segmentation using a CNN.*

The box plots in Figure 5.9 paint a similar picture. The interquartile range of /w/ and /r/ clearly overlap for lip height and horizontal position suggesting that the two phonemes do not greatly differ across these two dimensions. The main difference for /r/ and /w/ seems to involve the width and vertical position of the lips. A difference involving the vertical position of the lips may be suggestive of labiodentalisation: the bottom lip moves up towards the top teeth.

A generalised linear mixed-effects analysis was performed to predict the probability that a token is a /w/ based on the four automatic measures: width, height, vertical position and horizontal position of the lips. All four measures were z-scored to improve model fit and to

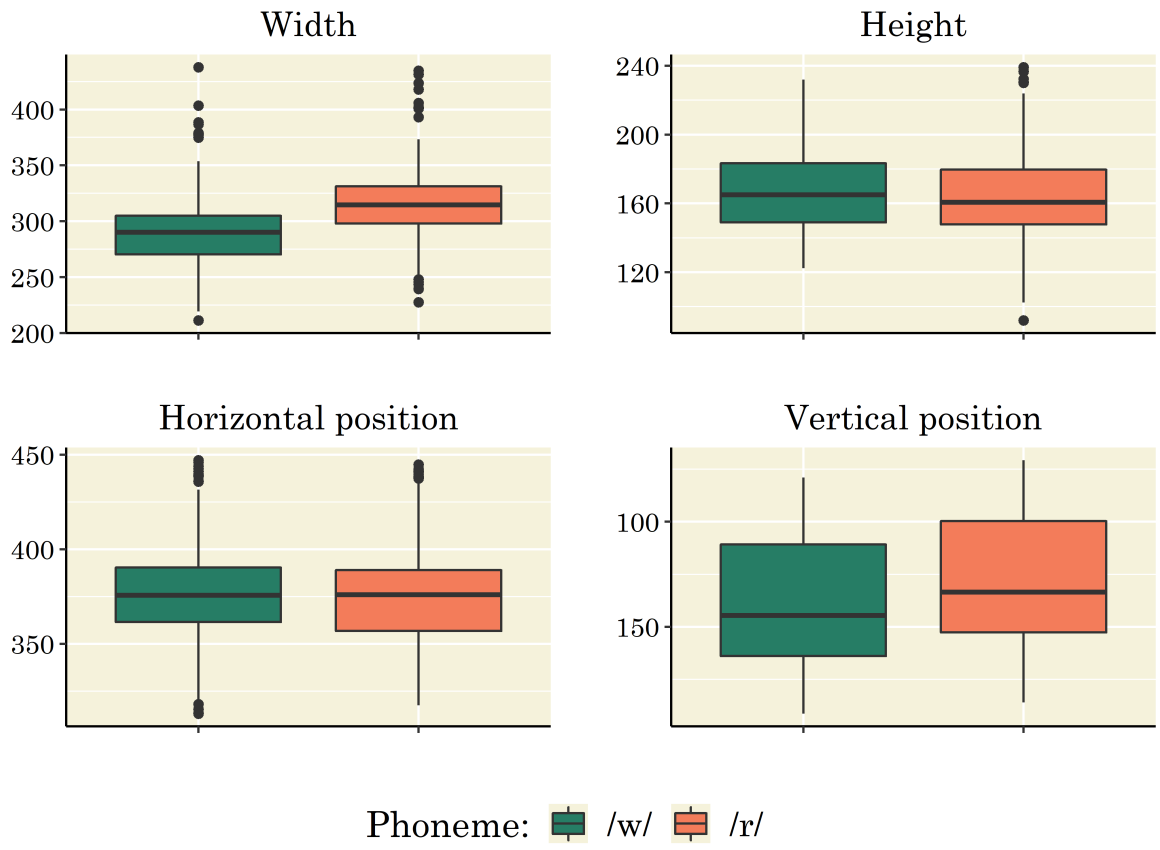



**Figure 5.9:** *Box plots presenting the lip dimensions (in pixels) of* /w/ *and* /r/ *acquired automatically from semantic segmentation using a CNN. The y-axis has been reversed for the vertical lip position measure to reflect the fact that lower values correspond to a higher lip position.*

allow us to measure the relative impact of all four variables on predicting the lip postures of /w/ and /r/. The random structure included by-Speaker random intercepts. The addition of by-Vowel random intercepts resulted in a singular fit and were thus removed. Likelihood ratio tests revealed that horizontal lip position was the only measure which failed to reach significance ($\chi^2(1) = 2.60, p = 0.11$). The three other dimensions were statistically significant predictors (Width: $\chi^2(1) = 159.93, p < .001$; Height: $\chi^2(1) = 25.75, p < .001$; Vertical Position: $\chi^2(1) = 80.06, p < .001$). We present the model output in Table 5.11, which shows that for an average speaker, the log-odds of observing a /w/ token are 4.45 higher when lip width decreases, 1.68 higher when lip height increases, and 7.27 higher when the lip position is low.[3] Although these three dimensions are statistically significant, a comparison of their $t$ values indicates that width and vertical position are the strongest predictors of phoneme category.

| Predictor | Estimate (log-odds) | Std. Error | $t$ value | $p$ value |
|---|---|---|---|---|
| (Intercept) | -0.08 | 1.67 | -0.05 | 0.97 |
| Width | -4.45 | 0.51 | -8.74 | $< .001^{***}$ |
| Height | 1.68 | 0.36 | 4.62 | $< .001^{***}$ |
| Horizontal Position | 1.18 | 0.75 | 1.58 | 0.12 |
| Vertical Position | 7.27 | 1.09 | 6.66 | $< .001^{***}$ |

*Phoneme ~ Width_z + Height_z + Horizontal Position_z + Vertical Position_z* + (1|*Speaker*)

**Table 5.11:** *Output of a generalised linear mixed-effects model predicting the probability a token is a /w/ according to the lip dimensions acquired automatically from semantic segmentation using a CNN. Lip dimensions were z-scored.*

We present the mean values observed for the three significant predictors (width, height and vertical position of the lips) per speaker for /r/ and /w/ in Figure 5.10, Figure 5.11 and Figure 5.12, respectively. We find the most robust difference between /r/ and /w/ to occur with regards to lip width and vertical position. Lip width is smaller on average for /w/ than for /r/ in 22 of the 23 speakers. Similarly, vertical lip position is higher for /r/ than for /w/ again in 22 of the 23 speakers. Lip height seems to be a less robust measure: 13 of the speakers have a

[3] A positive estimate for vertical position corresponds to a lower lip position.

larger lip height for /w/ than for /r/ on average.

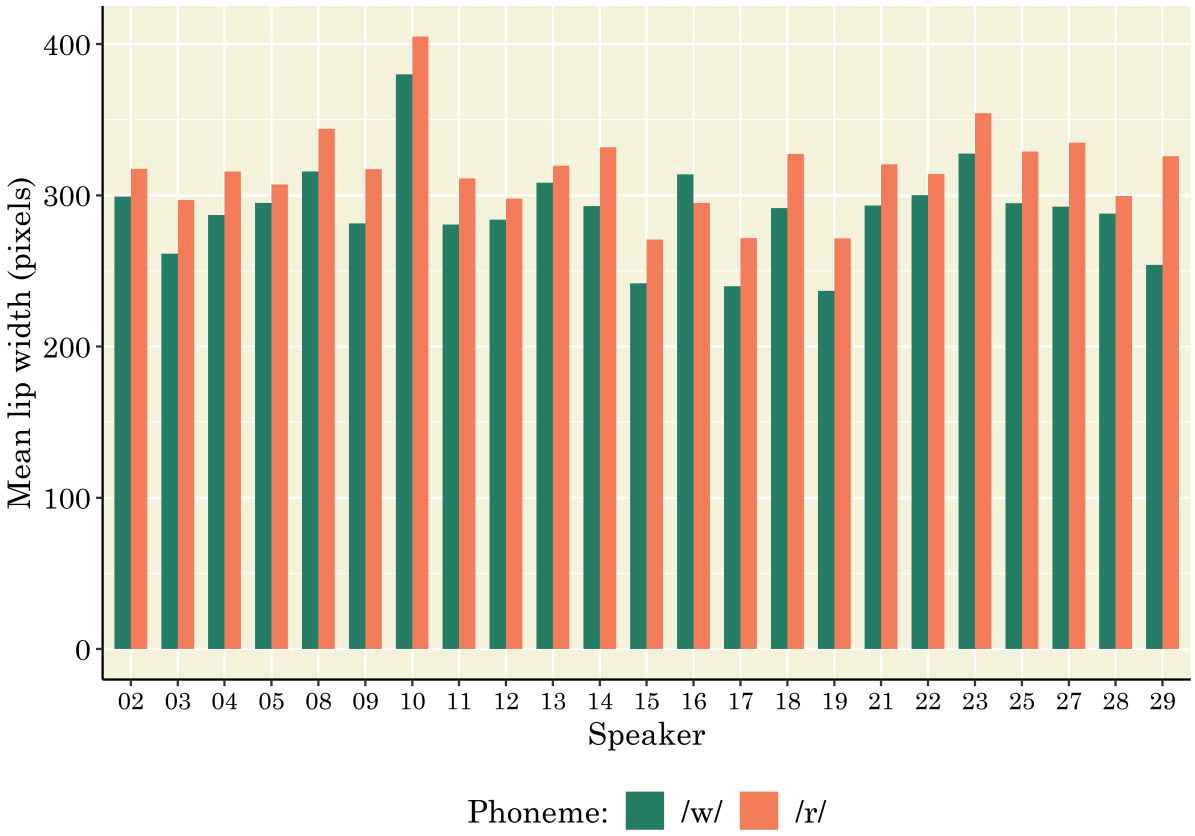



**Figure 5.10:** *Mean lip width (in pixels) of* /w/ *and* /r/ *in 23 speakers acquired via automatic semantic segmentation with a CNN.*

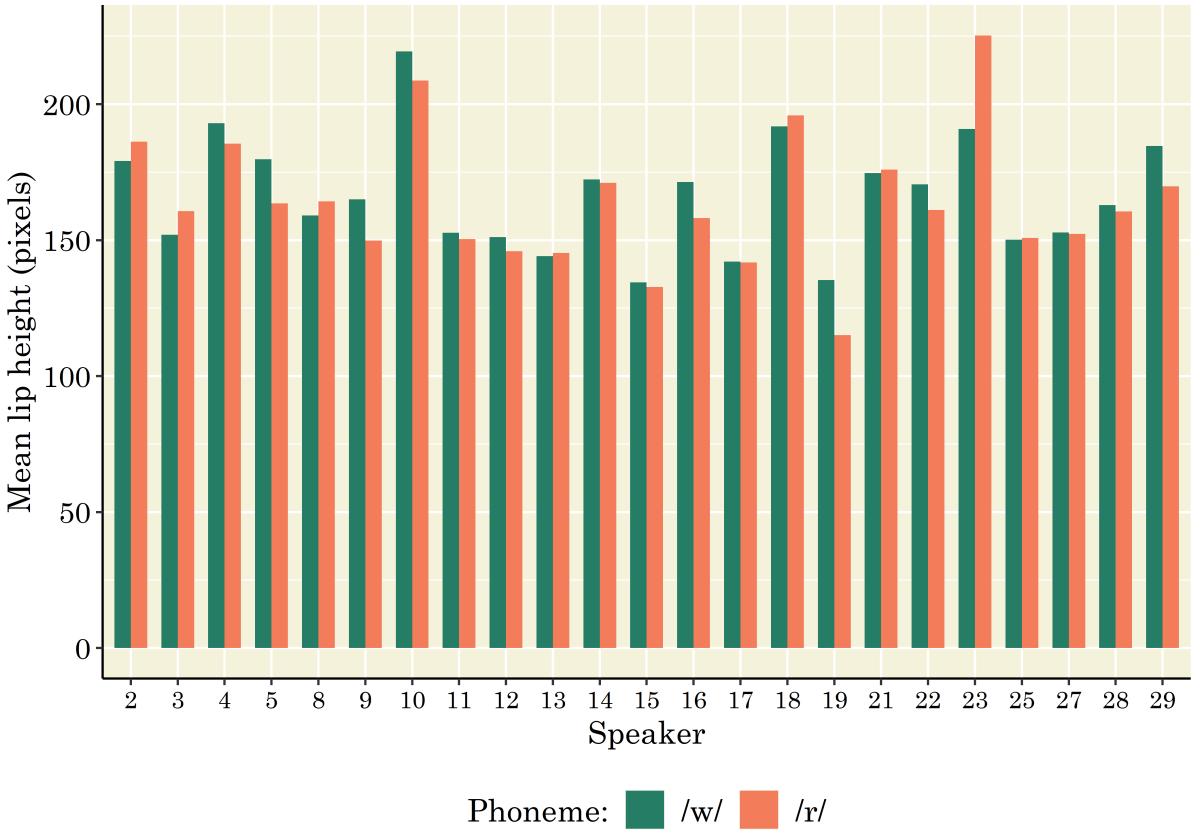



**Figure 5.11:** *Mean lip height (in pixels) of* /w/ *and* /r/ *in 23 speakers acquired via automatic semantic segmentation with a CNN.*

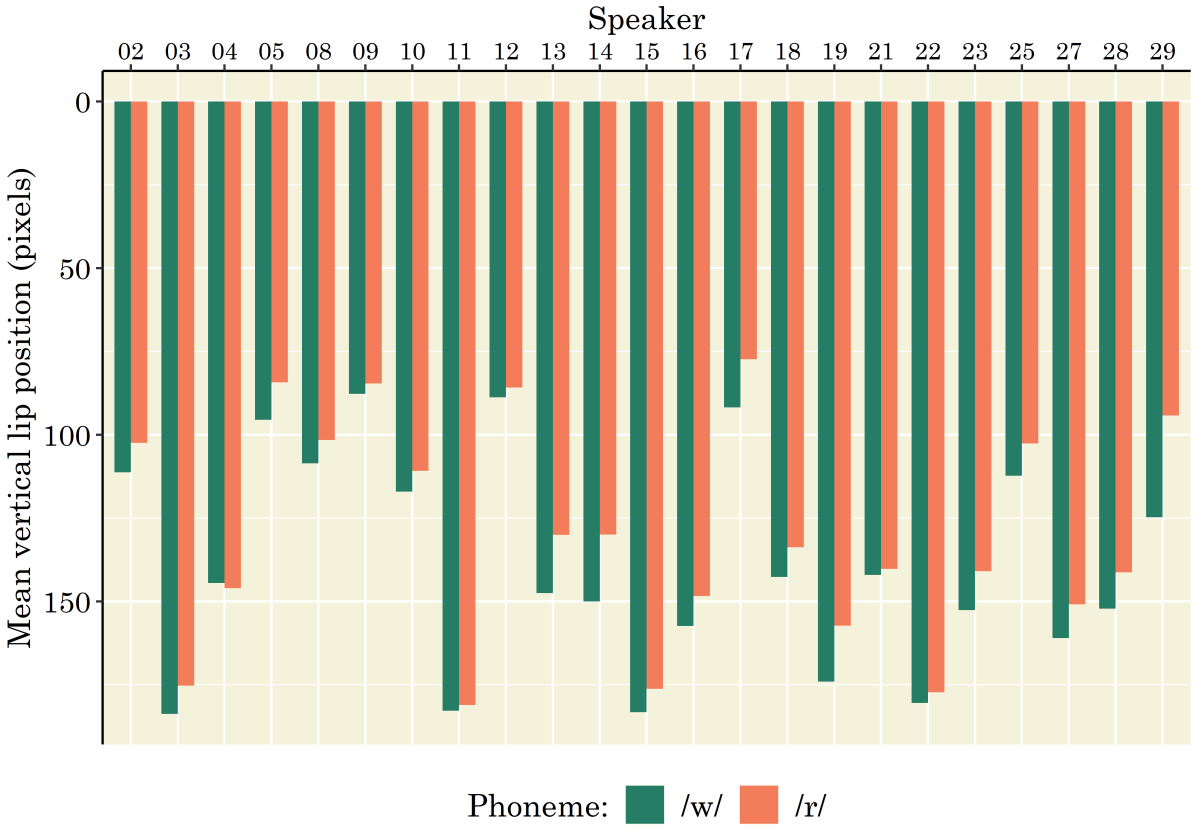



**Figure 5.12:** *Mean vertical lip position (in pixels) of* /w/ *and* /r/ *in 23 speakers acquired via automatic semantic segmentation with a CNN. The y-axis has been reversed to reflect the fact that lower values correspond to a higher lip position.*

### 5.3.3 *Automatic classification of /w/ and /r/ using a deep convolutional neural network*

In a separate analysis, another CNN was trained to automatically learn the difference between /w/ and /r/ from 414 raw images of the lips (without segmentation of the lip area). As discussed in Section 5.2.5, two types of model validation were applied: 10-fold cross validation and leave-one-out validation. 10-fold cross validation splits the dataset randomly into 10 equal subsets, 9 of which are used for training, while the remaining subset is set aside for testing. This process is repeated 10 times with each of the 10 subsets used for testing and the results are averaged to produce one single model estimation. The CNN achieved 99.52% mean correct classification of /r/ and /w/ tokens with a standard deviation of 1.02%. This very high model accuracy suggests that the front lip images for /r/ and /w/ differ. To ensure that the model used linguistically relevant information i.e., the lips, to classify /r/ from /w/, an occlusion analysis was performed. Figure 5.13 presents example heatmaps resulting from this occlusion analysis. Red regions in the heatmaps highlight the most relevant areas for the classification, while regions in blue (or those with no overlaid colour) show parts of the image whose influence on the classification is small to negligible. Visualising the heatmaps indicated that much more often than not, it is the lips that are highlighted. We can thus conclude with a reasonable degree of certainty that the lip configurations for /r/ and /w/ have sufficiently discriminant features which allow the programme to distinguish between them.

The second type of validation technique, leave-one-out validation, allowed us to challenge the generalisation ability of the models. Given that the dataset split is random in the previous 10-fold cross validation technique, data from all speakers is present in both training and test sets. This means that the models may have relied on speaker-specific information to distinguish between /w/ and /r/, rather than differences occurring across speakers. The leave-one-out validation procedure avoids this problem by leaving out a speaker's whole dataset for the testing stage. Training is therefore carried out with data from the remaining 22 participants. With this more demanding procedure, mean correct classification was 92.27 ±14.86%. Model accuracy varied from one speaker to the next, ranging from 50% to 100%. Model accuracy per

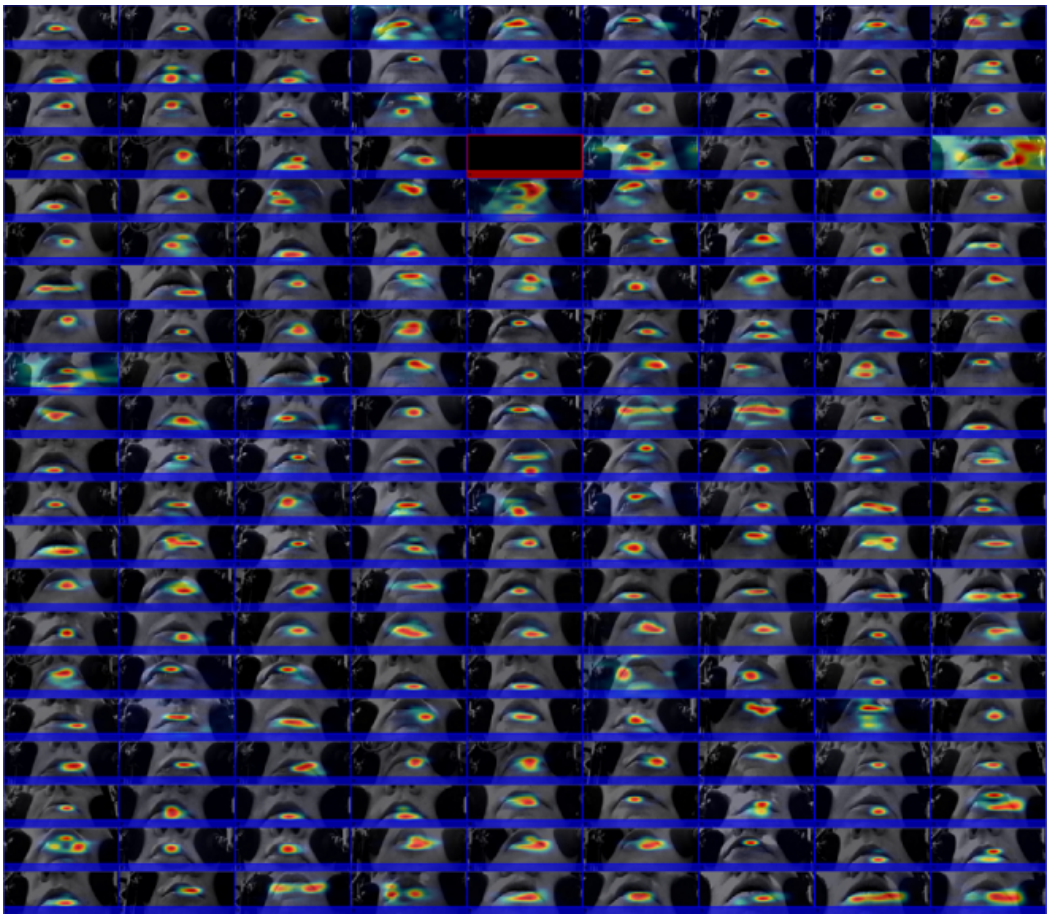


**Figure 5.13:** *Resulting heatmaps from occlusion analysis of a CNN trained to automatically classify* /w/ *and* /r/ *from 414 front lip images. The image with a red frame shows the only misclassified item in this batch.*

speaker is presented in Figure 5.14, which suggests that speakers 04, 13, 17 and 18 achieved the lowest accuracy scores. Again, occlusion analysis was performed, and for the speakers whose model accuracy score is high, the salient regions of interest for the models were, again, the lips. Visualisation of the data may give us some indication as to why productions in speakers were misclassified. The camera angle obscured the top lip to a certain extent in speakers 13 and 17, which could account for their low accuracy scores.

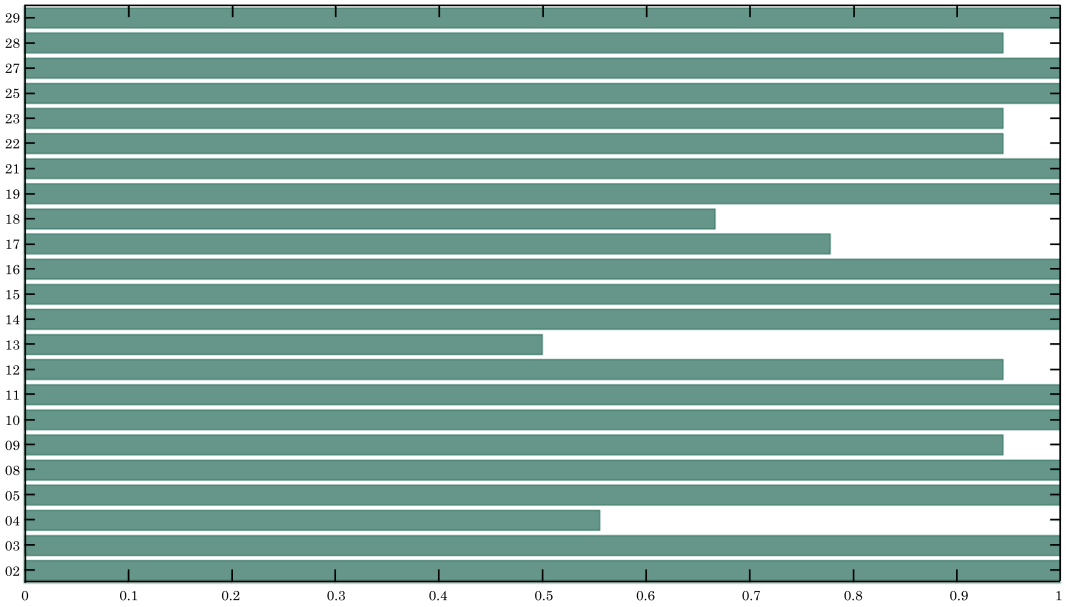



**Figure 5.14:** *Model accuracy per speaker of the automatic classification of* /w/ *and* /r/ *from front lip images using a CNN with a leave-one-out validation procedure.*

Figure 5.15 presents typical front view images of /r/ and /w/ from 12 speakers. Images were taken from productions of /r/ and /w/ followed by the FLEECE vowel from the words *weed* and *reed*, in order to avoid coarticulation with a following rounded vowel. Each subject's left-hand image corresponds to their lip posture for /w/. To facilitate comparisons between the lip postures for /w/ and /r/, we repeat here the two types of labialisation we defined in

Chapter 3:

> Horizontal labialisation: associated with back vowels, the lips are pouted by drawing the lip corners together to form a small, rounded opening.
>
> Vertical labialisation: associated with front vowels, the lips come together by raising the bottom lip and closing the jaw, resulting in a small, slit-like opening.

Generally speaking, the lip configurations are visibly different for /w/ and /r/ in Figure 5.15. Lip width is notably smaller for /w/ than for /r/, which is reflected in the results from both our manual and automatic lip measures. We notice that the lip corners are brought together for /w/ shortening the width of the lips, which is indicative of horizontal labialisation. As described in Section 3.1, horizontal compression of the lip corners is largely achieved by the contraction of the orbicularis oris muscle, which results in pronounced wrinkling of the labial skin (Folkins, 1978). Increased lip wrinkling is often apparent in the images of /w/ presented in Figure 5.15. The lip opening for /w/ is generally round and small, which is another feature of horizontal labialisation.

In contrast, in the images of /r/ in Figure 5.15 wrinkling of the labial skin is generally absent or much less noticeable than in /w/ and the lip opening tends to be more slit-like than round. We note that speaker 04 is the only speaker presented whose lip configurations for /r/ and /w/ appear to be somewhat similar: both have a small circular lip opening with a certain degree of wrinkling of the lip surface. Incidentally, this speaker had the smallest lip width for /r/ according to our manual lip measurements (as presented in Figure 5.8, p. 189). He also achieved one of the lowest accuracy scores in the classification of /r/ and /w/ by the CNN. Speaker 18, who also achieved low model accuracy, also produced /r/ with visible wrinkling of the lip surface, particularly in the context of the FLEECE vowel. Again according to our manual lip measurements, after speaker 4, speaker 18 had the smallest lip width on average for /r/ (as presented in Figure 5.8).

Finally, we observe from Figure 5.15 that the bottom lip is generally raised for /r/ in comparison to that of /w/. Automatic segmentation via a CNN allowed us to measure the

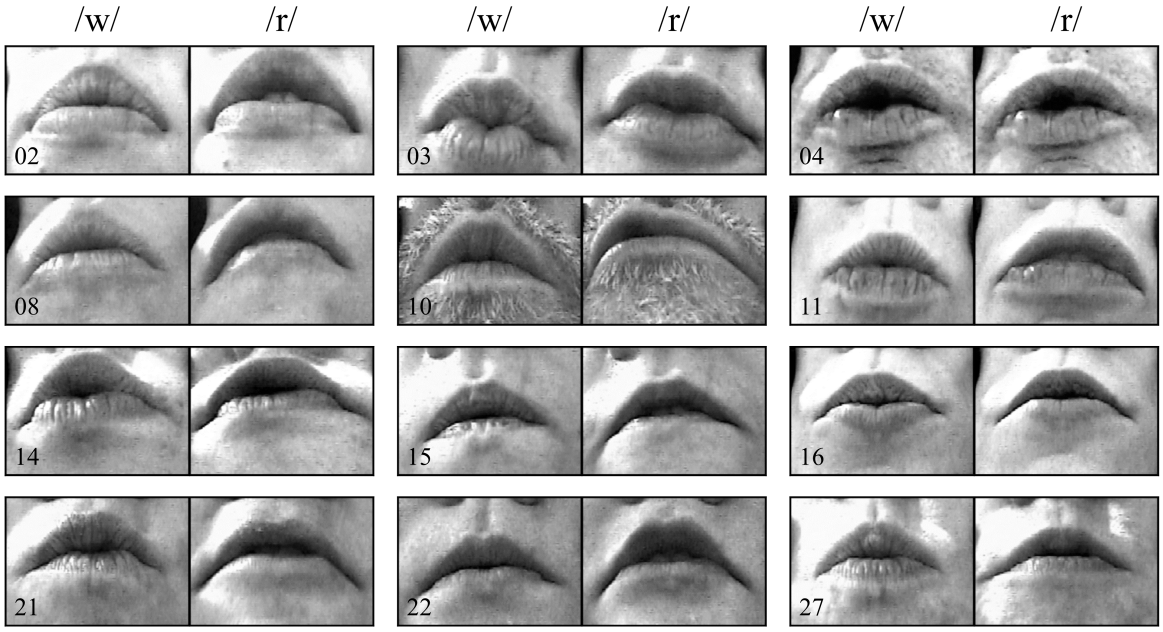



**Figure 5.15:** *Front view lip images of* /w/ *(left image) and* /r/ *(right image) from 12 speakers. Images depict maximal lip protrusion for* /w/ *and* /r/ *followed by the* FLEECE *vowel in the words* weed *and* reed.

vertical position of the lips and our statistical analysis indeed suggests that /r/ has a higher lip position than /w/. Both the raising of the bottom lip and the slit-like opening of the lips for /r/ seem to suggest that it is generally produced with vertical labialisation. In some speakers, the upper front teeth are visible during their /r/ production (e.g., speakers 02, 08, 11, 15, 21 and 22 presented in Figure 5.15). We notice that in the cases where the top teeth are visible, the inside of the bottom lip appears to be in close proximity to the front surface of the incisors, which is suggestive of an endolabial-dental articulation, as discussed in Section 5.1.

### 5.3.4 *Summary of results*

We have compared the acoustics and the lip postures of /w/ with those found in lingual productions of /r/ in Anglo-English and have shown that they differ in both respects. With regards to acoustics, F3 is around 800 Hz lower and F2 around 500 Hz higher for /r/ than /w/ on average in female speakers. However, /r/ and /w/ do not significantly differ in their respective

frequencies of F1. A visual inspection of front lip images indicates that the labialisation for /r/ and /w/ differs. Indeed, a CNN can tell a /w/ from a /r/ simply by 'looking' at the images of the lower part of a speaker's face without prior segmentation. Occlusion analysis revealed that it is the lips that the model relies on to categorise each image as /r/ or /w/. Measurements of the lip dimensions acquired both manually and automatically reflect the visible difference in lip posture between /r/ and /w/. Manual measures of lip protrusion, width and height revealed that the best predictor of phoneme is the width of the lips. In all 23 speakers, the lips are less wide for /w/ than they are for /r/. Lip protrusion and height were not statistically significant. Automatic measures from semantic segmentation using a CNN allowed us to consider the position of the lips as well as their height and width dimensions. Although width, height and vertical lip position were significant predictors of phoneme, the most robust indicators seem to be lip width (mirroring the results from manual lip measures) and vertical position. In 22/23 speakers, the lips are wider and higher on average for /r/ than they are for /w/. These differences seem to indicate that labialisation in /w/ is implemented via horizontal labialisation, whereas /r/ involves vertical labialisation. Finally, vertical labialisation, involving the raising of the bottom lip, seems to result in an approximation of the bottom lip with the top incisors, which is suggestive of an endolabial-dental articulation.

## 5.4 Discussion

### *5.4.1 Accounting for an /r/-typical labial gesture in Anglo-English*

The results from this study support Hypothesis 6: /r/ has a specific lip posture which differs from that of /w/ in Anglo-English. Quantitative analysis of both manual and automatic measures of the lip dimensions has revealed that what distinguishes the lip postures for /r/ and /w/ is predominantly the width (i.e., lip corner to corner) of the interlabial space and the vertical position of the lips. Statistical analysis of manual lip measures indicate that neither lip protrusion nor lip height are significant predictors of phoneme category, /r/ or /w/, although both dimensions are higher on average for /w/. Statistical analysis of lip measures acquired

automatically using a CNN paints a slightly different picture in that lip height is a significant predictor of phoneme category. However, when we consider each speaker's mean values separately, we find that the increased lip height predicted for /w/ only occurs in 13 of the 23 speakers. We suggest therefore that the strongest predictors of phoneme category are lip width and vertical lip position: 22 speakers have wider and higher lips on average for /r/ than for /w/. We conclude that the lip posture for /w/ involves horizontal labialisation, while /r/ is generally produced with vertical labialisation. /w/ is labialised horizontally because the lip corners are brought together towards the centre, forming a round shaped opening between the lips. For /r/, rather than bringing the lip corners to the centre, they are brought together vertically, creating an elliptical shaped lip opening.

Our results therefore indicate that the lip postures for /r/ and /w/ are phoneme specific in Anglo-English. Although it is well-established that lip rounding lowers formant frequencies by decreasing the size of the lip area and increasing the length of the vocal tract (e.g., Stevens, 1998; Vaissière, 2007), the exact acoustic consequences of the different lip postures we have described for /r/ and /w/ have yet to be accounted for. While the main acoustic correlate of /r/ is generally associated with a low F3, which is in close proximity to F2, the labio-velar approximant /w/ is characterised by a high F3 and a low F2 (as discussed in Section 5.1.1). Our acoustic analysis indeed found significant differences in F2 and F3 between /r/ and /w/. A connection may be made between the lip postures for /r/ and /w/ and those found in front and back vowels. As discussed in Chapter 3, front and back vowels are not produced with the same degrees of lip rounding. Back vowels are produced with horizontal labialisation (aka close lip rounding), resulting in a small lip area. Front vowels, on the other hand, are produced with 'less' lip rounding (Wood, 1986) i.e., with vertical labialisation. For back vowels, Stevens (1998) noted that in the case of a backed tongue position, the condition of minimum F2 is achieved only if the lips are rounded and a narrow opening is formed. For front vowels, it has been proposed that the lips are not as closely rounded as back vowels in order to maintain the proximity of F3 to F2. Catford (1977) explained that front vowels are usually 'exolabial' (Catford's equivalent to our vertical labialisation label), in order to avoid over-lowering the

second formant and hence preserve their front quality. Wood (1986) took this proposition one step further by suggesting that the difference in lip postures between front and back vowels is a linguistic universal as it is always the case that [y] is less rounded than [u]. We suggest then that by limiting the use of horizontal labialisation for /r/, Anglo-English speakers avoid over-lowering the second formant, thus conserving the proximity between the second and third formant for /r/ and ensuring a maximal perceptual contrast between /r/ and /w/.

Somewhat unexpected differences have been observed in the perception of approximants between American and Anglo-English listeners. In Dalcher et al. (2008), American and English participants judged whether copy-synthesised sounds with manually adjusted formant values were more like /r/ or /w/. A significant difference was observed for a stimulus which had a third formant typical of /r/ and second formant typical of /w/. American speakers identified this stimulus as /r/ 90% of the time, while Anglo-English speakers only identified it as /r/ 59% of the time. Dalcher et al. (2008) argued that the reason for such a disparity may be due to Anglo-English speakers being exposed to labiodental variants without a canonically low F3, unlike American English speakers. As a consequence, they speculated that F3 alone is no longer a sufficient cue to distinguish /r/ from /w/ in Anglo-English and that the F2 boundary between /r/ and /w/ may have become sharper in Anglo-English speakers. The fact that the vast majority of the Anglo-English speakers presented in this study use a lip configuration that potentially prevents them from over-lowering F2 (i.e., with vertical labialisation) seems to support Dalcher et al. (2008)'s hypothesis. Although all our speakers had an observable tongue body gesture with low F3 values typical of /r/, given the pressure to differentiate /r/ and /w/ beyond F3 due to exposure to high-F3 variants, Anglo-English speakers may find themselves in a delicate articulatory balancing act, having to make trade-offs between keeping F3 low without over-lowering F2. As F2 is less of a concern in Englishes with less exposure to high-F3 /r/ variants, we predict that American English speakers would be freer to use more variable, more [w]-like lip postures for /r/ in order to enhance the salience of /r/. The findings from a very recent study on American English support this hypothesis. Labial postures presented in B. J. Smith et al. (2019) were much more variable across speakers, with more instances of

horizontal labialisation reported for /r/ than in our Anglo-English data.

The lip posture we have described for /r/ in Anglo-English, which we suggest is used by speakers with high exposure to labiodental /r/ in order to enhance F3 lowering but avoid over-lowering F2, seems to rather ironically share similar features to labiodental articulations. In order to protrude the lips without horizontal labialisation, the lower lip is raised towards the top front teeth, which was described as vertical compression of the lip corners by Catford (1977). The inner surface of the lower lip is thus in close proximity with or perhaps even touching the upper front teeth, resembling an endo-labiodental articulation (as described in Section 5.1.1). In contrast, horizontal labialisation, in which the lip corners of the mouth are drawn together away from the front teeth along the occlusal plane, makes contact between the lips and front teeth almost impossible. We speculate then that the typical lip posture accompanying lingual productions of /r/ in Anglo-English results in the approximation of the lower lip and the top teeth, or labiodentalisation. Labiodental variants may thus continue to emerge if the labial gesture takes precedence over the lingual one, as suggested by Docherty and Foulkes (2001), particularly if the labial gesture is visually prominent. Indeed, in some speakers, the proximity of the bottom lip and the upper front teeth is clearly visible during their /r/ production. As a result, like Dalcher et al. (2008), we also predict increased use of labiodental /r/ in Anglo-English in the future.

The present study therefore confirms the assumption that the post-alveolar approximant /r/ is generally produced with a labiodental-like lip posture in Anglo-English, which may therefore motivate the change towards labiodental articulations of /r/ (e.g., [ʋ]) resulting from the loss of the lingual gesture, as proposed by Docherty and Foulkes (2001), Foulkes and Docherty (2000) and Jones (1972). We have suggested that this /r/-typical labial gesture ensures a maximal acoustic contrast between /r/ and /w/. Vertical labialisation allows speakers who produce /r/ with an observable tongue body gesture to enhance the low frequency of F3 by lengthening the cavity in front of the lingual constriction all the while maintaining a small distance between F3 and F2 by avoiding horizontal labialisation. However, an alternative explanation for the observed difference in labial configurations between /r/ and /w/ in Anglo-English could

be that by using distinctive articulatory cues, speakers are able to enhance the perceptual contrast between the two sounds in the visual domain. Indeed, speech has been shown to be visually optimised in cases where pressure to maintain a phonological contrast is high. For example, Havenhill and Do (2018) observed that in American English, the visual lip rounding cue enhances perception of the /ɑ/-/ɔ/ contrast, and Traunmüller and Öhrström (2007) found that in Swedish, listeners rely on visual cues in the perception of /i/-/y/. Indeed, Docherty and Foulkes (2001) proposed that the loss of the lingual gesture in Anglo-English /r/ may be due to the heavy visual prominence of the lips. We have shown in this study that a Convolutional Neural Network can distinguish between /r/ and /w/ with very high levels of accuracy just by 'looking at' images of the lips. Although the human brain has long served as a source of inspiration for machine learning and the best algorithms today for learning structure in data are artificial neural networks (Fong, Scheirer, & Cox, 2018), we do not know if the difference in lip postures between /r/ and /w/ is perceptually salient to human listeners. Our next task will therefore be to assess to what extent the visual cue of the lips influences the perception of the /r/-/w/ contrast in humans, which will be considered in depth in the next part of the thesis.

### *5.4.2 Methodological implications*

Finally, on a methodological level, we have used techniques from deep learning to not only train models to learn articulatory differences from raw lip images, but also to automatically segment and measure the lips. That Convolutional Neural Networks learn their own representations from the data constitutes a promising research avenue for future phonetic studies. We see no reason why analyses with CNNs may not be extended to any relatively large image-based dataset, such as those containing Ultrasound Tongue Imaging, spectrograms and fundamental frequency curves to name a few. This study has shown that it may be possible to partly overcome the 'black box' problem and make what DNNs learn from the data more explicit using occlusion analysis. We have illustrated how the visualisation of heatmaps not only makes neural networks' decisions more interpretable, but can also draw researchers' attention to potential biases in their studies. We were able to show that the models drew on linguistically-relevant

information within the images (i.e., the lip area) to classify tokens as /w/ or /r/. Had this not been the case, the high accuracy obtained by the model would have been very hard to interpret.

Semantic segmentation using a CNN was able to accurately segment the lip area from the rest of the image and provided us with measurements of the dimensions and position of the lips, despite the quality of the lip images being rather poor due to adverse recording conditions. This approach was less time consuming and is more reproducible than taking measurements of the lips by hand. Although we have presented results from static data, a logical extension will be to train models with whole lip videos rather than selected frames, which we are currently working on implementing.

## 5.5 CHAPTER CONCLUSION

We have presented articulatory evidence to show that lingual productions of /r/ are accompanied by a specific labial gesture which is distinct from that of /w/ in Anglo-English. While /w/ is produced with horizontal labialisation, /r/ generally involves vertical labialisation. We have related the development of this /r/-typical lip posture to Anglo-English speakers' increased exposure to labiodental variants of /r/ and to the ensuing pressure to maintain a perceptual contrast between /r/ and /w/. We suggest that vertical labialisation enables speakers with an observable tongue body gesture to maintain a low F3 without over-lowering F2. Over-lowering of F2 could cause perceptual uncertainty as the acoustic cue that distinguishes a high-F3 (non-lingual) /r/ from /w/ may now be F2 (Dalcher et al., 2008). In Englishes where high-F3 variants are not reported, the frequency of F3 remains the most prominent acoustic cue for /r/ (Dalcher et al., 2008), which we predict allows speakers more freedom to vary the accompanying lip gesture for /r/, which may account for the differences observed between the labial gesture in the present study and that presented in B. J. Smith et al. (2019) in American English. Finally, in avoiding over-lowering F2 due to increased exposure to labiodental /r/, the lip posture in speakers who still have an observable tongue body gesture has perhaps inadvertently become more labiodental. Following Dalcher et al. (2008), we also predict a further increase in labiodentalisation in Anglo-English /r/. The cue for /r/ in Anglo-English

may continue to shift to F2 to such an extent that speakers will attend less to F3, provoking them to retain the labiodental component of their articulation at the expense of the lingual one. The change towards exclusively labial articulations of /ɹ/ may progress even more rapidly if the labial gesture is particularly visually prominent. We will therefore turn our attention to the impact of the visual cue of the lips on the perception of /ɹ/ in the next part of this thesis.

*Part* III

# Perception of Anglo-English /r/

# Audio-visual perception of Anglo-English /r/

# 6

## 6.1 Introduction

The results from Experiment 2 indicated that the phonetic implementation of labialisation for /r/ differs from that of /w/ in Anglo-English. We concluded that lingual /r/ may have developed its own specific lip posture due to increased exposure to non-lingual variant and to the resulting pressure to maintain a perceptual contrast between /r/ and /w/. Up to now, we have related the development of this labial posture to the *auditory* perception of /r/: vertical labialisation allows speakers who still produce /r/ with a lingual constriction to maintain a low F3 without over-lowering F2. However, this /r/ specific labial posture may have also evolved in order to enhance the perceptual salience of /r/ in the *visual* domain as well as the auditory one. Indeed, as we observed in Chapter 1, we suggested that languages may have evolved and may continue to evolve to ensure that sound contrasts which are difficult to hear are easy to see. The evolution of separate visemes could therefore help disambiguate /r/ and /w/ when perceived visually.

### 6.1.1 *Aims and predictions*

In this experiment, we aim to assess to what extent the labial gesture for /r/ in Anglo-English is perceptually salient by considering the perception of /r/ and /w/ in English subjects in the following presentation modalities: auditory-only, visual-only, congruous audio-visual and incongruous audio-visual. If the labial gesture of /r/ is visually prominent, we expect the visual cue of the lips to enhance the auditory perception of the /r/-/w/ contrast. English subjects should thus be able to discriminate /r/ from /w/ better in the audio-visual modality than in the auditory-only one. As discussed in Chapter 1, previous research has generally shown that subjects perform less well in visual-only than auditory-only perception of speech, i.e., the visual cues are less informative than the auditory ones. However, if a specific visual cue for /r/ has evolved in order to ensure that the /r/-/w/ contrast, which is increasingly difficult to hear due to growing exposure to non-lingual productions of /r/, remains visually perceptible, /r/ may be best discriminated from /w/ in visual-only as opposed to auditory-only perception. We aim to further test this prediction by presenting English subjects with incongruous audio-visual stimuli for /r/ and /w/, i.e., visual /r/ will be paired with auditory /w/ and visual /w/ will be paired with auditory /r/. As other studies examining incongruous audio-visual speech perception have pointed out, visual capture may be anticipated if the visual cues for /r/ and /w/ are unambiguous (i.e, are visemic) and are more perceptually salient than their auditory cues. Our predictions concerning the perception of the /r/-/w/ contrast across different modalities can therefore be summarised as follows:

- auditory-only < audio-visual
- auditory-only < visual-only
- auditory responses < visual responses in incongruous audio-visual perception

By including a control stimulus /l/, which is not produced with labialisation, we will be able to test whether /r/ and /w/ are distinguishable from a non-labial segment. If /r/ and /w/ are both produced with labialisation, subjects should be able to distinguish /r/ from /l/ and /w/

from /l/ in the visual-only modality, as well as in the auditory-only and the audio-visual ones. If the aforementioned arguments are valid, the following hypothesis may be derived:

**Hypothesis 7** Perceptual sensitivity to the /r/-/w/ contrast is enhanced by visual cues of the lips in Anglo-English.

### *6.1.2 Viseme mappings for /r/ and /w/ in the literature*

In the literature on visemes, as discussed in Section 1.1, the treatment of /r/ and /w/ varies. Bear et al. (2014) present the phoneme-to-viseme consonant mappings in English according to 15 previous studies. We will consider 6 of these 15 studies, which were selected because viseme classes were mapped according to the results from perception data in humans (and not in machines) and simply due to their accessibility. We also included a further follow up study by Walden and colleagues which was not included in Bear et al.'s original review. The seven studies considered here are the following:

1. Binnie, Jackson, and Montgomery (1976)
2. Fisher (1968)
3. Franks and Kimble (1972)
4. Kricos and Lesner (1982)
5. Woodward and Barber (1960)
6. Walden, Prosek, Montgomery, Scherr, and Jones (1977)
7. Walden, Erdman, Montgomery, and Schwartz (1981)

All seven studies investigated consonant visemes in American English but did not necessarily converge in their attribution of viseme classes to /r/ and /w/. Binnie et al. (1976), Fisher (1968), Franks and Kimble (1972), Kricos and Lesner (1982) and Woodward and Barber (1960) all considered the lip reading capabilities in normal hearing adults in varying numbers of subjects ranging from 12 to 275. Binnie et al. (1976) is the only study of the five which explicitly suggested

that there are two separate visemes for /r/ and /w/. The remaining four all indicated that they belong to the same viseme. Binnie et al. (1976) proposed that this particular inconsistency may be due to the more favourable viewing conditions in their study than in the others, which resulted in a larger number of viseme categories. However, even in the four studies which posited one distinct viseme for /r/ and /w/, we observe inconsistencies in their results. Kricos and Lesner (1982) presented subjects with visual productions from speakers who vary in ease of being lip read and found differing viseme mappings for each individual speaker. Woodward and Barber (1960) indicated that while the visual cues for /r/ and /w/ are very similar, /r/ and /f/ are similar but /w/ and /f/ are contrastive, which suggests that /r/ shares more labial properties with labiodentals than /w/. Fisher (1968) considered the viseme mapping of /r/ and /w/ to be 'directional' in that /r/ as a stimulus was significantly confused with /w/, but /w/ was not confused with /r/.

By pooling together the results from these studies, the inconsistencies we observe in the treatment of /r/ and /w/ suggest that in some cases, it may be possible to distinguish their visual cues in American English and that this may be linked to individual differences in production. Indeed, Walden et al. (1977) and Walden et al. (1981) both considered the lip reading capabilities in hearing-impaired subjects but presented different viseme mappings for /r/ and /w/. In Walden et al. (1977), hearing impaired subjects could distinguish /r/ from /w/ with over 75% accuracy from lip reading alone, making the authors posit two distinct visemes for /r/ and /w/. In contrast, Walden et al. (1981) could not justify attributing separate viseme mappings for /r/ and /w/ from their data. They suggested that the reason for the disparity in the results between the two studies may lie in variation across the speakers presented as stimuli. In Walden et al. (1977), the speaker had undergone training and appeared easier to lip read than the untrained speaker used in Walden et al. (1981).

The seven studies reviewed here are undeniably dated. Modern studies which present phoneme-to-viseme mappings assess automatic speech recognition in machines and visual perception in human listeners is therefore less of the focus than it was forty years ago. Moreover, as far as we are aware, phoneme-to-viseme mappings for Anglo-English based on the visual

perception of speech in humans do not currently exist. It is hoped that the present study will provide more insights into possible viseme mappings for /ɹ/ and /w/ in Anglo-English.

### 6.1.3 *Quantifying visual enhancement*

A common way to quantify the benefit obtained from adding a visual signal to an auditory stimulus (i.e., to measure visual enhancement) is to calculate the difference between a listener's performance in audio-visual and auditory-only conditions expressed relative to the amount of possible improvement given the subject's auditory-only score (e.g., Grant & Seitz, 1998; Grant, Walden, & Seitz, 1998; Sommers et al., 2005; Sumby & Pollack, 1954; Van Engen, Xie, & Chandrasekaran, 2017). Visual enhancement may thus be calculated according to the following equation:

$$\text{visual enhancement} = \frac{(\text{audio-visual} - \text{auditory-only})}{(1 - \text{auditory-only})}$$

As Sommers et al. (2005) pointed out, this measure prevents the bias resulting from a simple difference score (i.e., auditory-only − audio-visual), as higher proportions in auditory-only performance would necessarily lead to lower enhancement values. However, the visual enhancement measure requires the proportion of correct responses in the auditory-only modality to be less than one (i.e., subjects must not have perfect accuracy) because we cannot divide by zero. There must therefore be room for improvement from the auditory-only modality to the audio-visual one (Van Engen et al., 2017). To prevent subjects from reaching ceiling in their auditory-only responses, researchers tend to add noise to their auditory stimuli, allowing them to calculate visual enhancement with this particular measure.

## 6.2 Methodology

The perception of /ɹ/ was assessed using a two-alternative forced choice identification (2AFC) task in which participants were presented with a target word beginning with [ɹ], [w] or [l]. Words beginning with [l] were included as controls as [l] is not produced with labialisation. After being presented with the target word in isolation, subjects were asked to identify the

consonant they perceived from two word options beginning with 'r' or 'w', 'l' or 'r', and 'l' or 'w'. Stimuli were masked with noise and were presented in the following four modalities:

1. auditory-only (AO)
2. visual-only (VO)
3. congruous audio-visual (AVc)
4. incongruous audio-visual (AVi)

Auditory stimuli were embedded in noise in order to prevent participants from attaining perfect identification scores in the auditory-only modality and to allow them room for improvement with the addition of visual cues in the audio-visual (AVc) modality. It was hoped that the inclusion of noise would allow us to implement the visual enhancement measure presented previously (in Section 6.1.3).

### *6.2.1 Participants*

40 native Anglo-English speakers (21F, 19M) aged between 18 and 73 (mean = 41.32 ± 17.92) took part in the perception study, which was conducted in North Yorkshire, England. Some participants (n=8) were recruited at the University of York, where ethical approval had been granted. Subjects at the university were undergraduates and were either financially compensated (£5) for their participation or gained class credit for linguistics courses. The remaining 32 participants were recruited by employing the 'friend of a friend' technique (Milroy & Gordon, 2008) with the author's existing connections in the area. Although these 32 subjects were offered monetary compensation, all of them chose to participate voluntarily. All participants self-identified as speaking with an English accent and we made sure that this was indeed the case by conversing with participants before recording them. Conversing with the subjects also allowed us to informally classify the participants' accents as rhotic or non-rhotic. One subject (11), who comes from the north west of England, had a rhotic accent. The remaining 39 subjects were non-rhotic.

Before participating, subjects signed an informed consent form (presented in Appendix B.4) and completed a background questionnaire (presented in Appendix B.5). Demographic information for all 40 participants is presented in Table 6.1. Although all participants spoke with an Anglo-English accent, three of them spent the majority of their childhood abroad. The dataset contains responses from two bilingual speakers: one English-French and one English-Tagalog. Table 6.1 also presents the foreign languages spoken by the participants beyond beginner level. Language proficiency is presented according to the Common European Framework of Reference for languages (Council of Europe, 2001). None of the participants reported to have any known speech or language disorders. One participant (21) had an uncorrected sight problem. Another subject (13) wore hearing aids and two subjects (10 & 34) reported experiencing occasional bouts of tinnitus. Six questions were included in the background questionnaire in order to assess the participants' hearing. Subjects were asked to judge to what extent their hearing suffered in typical listening scenarios using the following six questions, which are based on the ones found in online hearing tests[1]:

1. Do you feel like you have any hearing problems, which are not currently known or treated?
2. Do you sometimes find it challenging to have a conversation in quiet surroundings?
3. Do you find it difficult to understand speech on TV and radio?
4. Do you find it difficult to follow conversations at dinner parties?
5. Do you find yourself having to ask people to repeat themselves?
6. Do you find it hard to have a conversation on the phone?

For each of the six questions, participants were asked to select the most appropriate response from the following: 'always, often, sometimes, rarely, never'. Their responses were converted into a score out of 30, which is presented in Table 6.1. A 'never' response obtained 5 points,

[1] An audiologist was consulted who recommended this set of questions from the Widex website. However, she stressed that self-assessment could evidently not replace clinical evaluations of hearing, which could not be implemented here.

| Subject | Sex | Age | Origin | Languages | Hearing |
|---|---|---|---|---|---|
| 01 | F | 26 | north east | French B2; Italian B2 | 30 |
| 02 | M | 61 | north east | | 28 |
| 03 | F | 47 | south East | | 24 |
| 04 | M | 52 | north east | | 24 |
| 05 | F | 57 | abroad | | 30 |
| 06 | M | 52 | north east | German C2; French B1; Spanish B1 | 30 |
| 07 | F | 30 | north east | Italian B2 | 28 |
| 08 | F | 52 | south east | French C1; Italian B1 | 27 |
| 09 | M | 60 | north east | | 30 |
| 10 | M | 53 | north east | French B1 | 26 |
| 11 | M | 62 | north west | | 30 |
| 12 | M | 73 | south east | | 28 |
| 13 | F | 73 | north west | | 21 |
| 14 | M | 56 | north east | | 30 |
| 15 | F | 30 | north east | | 25 |
| 16 | M | 29 | north east | | 23 |
| 17 | F | 32 | north east | | 29 |
| 18 | M | 30 | north east | | 26 |
| 19 | M | 65 | midlands | | 23 |
| 20 | F | 56 | north east | | 22 |
| 21 | M | 64 | north east | | 28 |
| 22 | M | 25 | north east | | 12 |
| 23 | F | 47 | abroad | | 26 |
| 24 | M | 47 | north east | | 21 |
| 25 | M | 43 | south east | | 21 |
| 26 | F | 61 | south east | French B1 | 19 |
| 27 | F | 64 | north east | Polish B1 | 23 |
| 28 | M | 21 | south east | Spanish C1; Italian B2 | 24 |
| 29 | F | 19 | north west | Tamil B2 | 28 |
| 30 | F | 18 | midlands | French bilingual; German C1; Japanese B1 | 25 |
| 31 | M | 19 | north west | | 23 |
| 32 | F | 19 | north west | Spanish B2; German B1 | 25 |
| 33 | F | 19 | south east | | 26 |
| 34 | F | 26 | north east | | 24 |
| 35 | M | 21 | north east | | 16 |
| 36 | M | 21 | north east | | 22 |
| 37 | F | 18 | south west | | 23 |
| 38 | F | 20 | abroad | Tagalog bilingual | 24 |
| 39 | F | 55 | north east | | 26 |
| 40 | F | 30 | north east | | 30 |

**Table 6.1:** *Participant demographics from the perception experiment.*

while 'always' received 1 point. As such, a score of 30 indicates that a subject self-identifies as presenting no signs of hearing loss. Based on the questionnaire responses, no participants were eliminated from the study.

### 6.2.2 Stimuli

A list of monosyllabic minimal pairs contrasting /r/, /w/ and /l/ word-initially in CV(C)C mono-syllabic words was produced. To avoid labial coarticulation with the following vowel, the onsets occurred in the context of six non-rounded vowels from the following lexical set: FLEECE, KIT, DRESS, TRAP, FACE, PRICE. Similarly, the coda consonant of each minimal pair was never a labial. Each onset and vowel combination was assigned two items resulting in 36 words (3 onsets × 6 vowels × 2 items). Test words are presented in Table 6.2. A list of the same number of minimal pairs contrasting /th/[2], /s/ and /h/ word-initially was also produced to act as fillers and controls, some of which contained rounded vowels and labial codas. A full list of these filler and control words is presented in Appendix B.1.

| Lexical set | /r/ | /w/ | /l/ |
|---|---|---|---|
| FLEECE | *reed* | *weed* | *lead* |
| | *reek* | *week* | *leek* |
| KIT | *rit* | *wit* | *lit* |
| | *rick* | *wick* | *lick* |
| DRESS | *red* | *wed* | *led* |
| | *rent* | *went* | *lent* |
| TRAP | *rack* | *whack* | *lack* |
| | *rag* | *wag* | *lag* |
| FACE | *rate* | *wait* | *late* |
| | *rake* | *wake* | *lake* |
| PRICE | *right* | *white* | *light* |
| | *rise* | *wise* | *lies* |

**Table 6.2:** *Test words comprising 36 monosyllabic* /r, w, l/ *word-initial minimal pairs grouped according to lexical set vowel.*

[2] /th/ will be used to indicate both voiceless /θ/ and voiced /ð/. The voiced interdental was used when minimal pairs with the voiceless interdental were not attested.

### Recording stimuli

One female 22-year-old native Anglo-English speaker was video-recorded producing the full word list in a sound-attenuated recording booth at the Université de Paris. Prior to recording, the speaker signed an informed consent form and completed a background questionnaire, which were very similar to the ones used for the production experiment presented in Appendices A.1 and A.2. Conversing with her beforehand allowed us to confirm that she audibly produced a post-alveolar approximant /ɹ/.[3] The speaker, who speaks French at an advanced level, had been living in France for just over one year and is an English teaching assistant at the university. Prior to moving to France, she had spent her whole life in the Midlands. She reported to have no known speech or hearing problems.

Audio and video recordings were made using a Zoom Q2HD Handy Video Recorder, which was chosen because it combines high quality audio with high definition video. Video was set to 1 280 × 720 resolution (in pixels) recording 59.94 frames per second. The video camera's built in condenser microphone was used to record the accompanying audio signal, which was digitised as a PCM stereo file with a 44 100 Hz sampling rate and 16-bit quantization. The resulting audio file was converted from stereo to mono during the post-processing stage in Praat by extracting the left channel.

Efforts were made to ensure that the video camera remained in a fixed position relative to the speaker's head. It was agreed that simply attaching the video camera to a tripod was not sufficient because we wanted to ensure that the lips remained in the same position throughout the recording session. Any movements made by the speaker would have changed the position of the lips within the shot. By keeping the lips in the same place in the video frame throughout the perception trials, we hoped to facilitate the participants' task of continuously watching the speaker's mouth. A stabilisation device was thus designed and produced using a bike helmet.[4] The video camera could be positioned directly in front of the speaker's lips by attaching it to

[3] We had planned on using Ultrasound Tongue Imaging to confirm that the speaker had an observable tongue body gesture for /ɹ/, but were hindered by university closures.

[4] Special thanks go to Emmanuel Ferragne and his brother for their ingenious creation!

the bike helmet via a flexible arm (recycled from a pop filter) and a handheld tripod. An image of video camera stabilisation is presented in Figure 6.1.


**Figure 6.1:** *The author demonstrating video camera stabilisation using a bike helmet, a flexible arm and a handheld tripod.*

The speaker was provided with the full word list in advance so that she could familiarise herself with the stimuli. We also made sure that there were no pronunciation ambiguities. For example, she was asked to pronounce the word *lead* with the FLEECE and not the DRESS vowel. Familiarising the speaker with the word list was also important because some of the words are evidently much less frequent than others, e.g., *rit* versus *red*. Once she was accustomed to the word list, the video camera stabilisation helmet was fitted. We ensured that the video camera remained in a constant position even when she moved her head by tightening the fastening underneath her chin. The video camera was then positioned directly in front of the speaker's mouth capturing the bottom half of her face from the nose to just under the chin.

Once the video camera was in position, we pressed the record button on the camera and left the recording running until the end of the session. The camera recorded directly onto

an SD card producing a file compressed with a MPEG-4 AVC/H.264 (MOV) video codec to Quicktime format. The speaker was instructed to keep a neutral expression and to speak clearly throughout. She was seated in front of a computer monitor showing the words approximately at eye level. The stimuli were presented in a slideshow using Microsoft PowerPoint®. Each slide contained one word from the word list, which was produced in isolation. The word list occurred in a semi-randomised order, as sequences of words with /r/ and /w/ onsets were prohibited. Slide transition timing was set to a 2-second advance. By giving the speaker the same amount of time to produce each word, it was hoped that she would produce words of a similar length, which would facilitate the creation of incongruous audio-visual trials. To accustom the speaker to the speed of presentation of the stimuli, the recording session began with a practice round containing 10 minimal pair /hVd/ words. Five tokens of each of the 36 test words and the 36 filler words were recorded. The speaker was provided with four 20-second breaks and the whole recording session lasted just under 15 minutes.

### *6.2.3 Generating perception trials*

Perception trials were created from the digital video of the Anglo-English speaker producing minimal pairs contrasting word-initial /r, w, l/ and word-initial /th, s, h/. The audio-visual recordings were edited using VirtualDub (A. Lee, 2000) to create perception trials in the following four modalities, which are presented schematically in Figure 6.2:

1. **Visual-only (VO)**: the audio track is replaced with noise and the original video track is maintained.

2. **Auditory-only (AO)**: the video track is replaced with a still image of the speaker's face and the audio track is embedded with noise.

3. **Congruous audio-visual (AVc)**: the audio track is embedded with noise and the original video track is maintained.[5]

[5]We had originally planned on using different tokens of the same word to produce congruous audio-visual trials but were unable to find tokens which matched closely enough in word length to create naturalistic materials.

4. **Incongruous audio-visual (AVi)**: auditory /r/ embedded with noise is dubbed over visual /w/ and vice versa, and auditory /s/ embedded with noise is dubbed over visual /th/ and vice versa.

Although the Anglo-English speaker produced five tokens of the 72 items in the word list, three tokens were used to generate trials for the perception experiment. The same token of each item was used to create auditory-only and visual-only trials. A different token was used to produce trials in the congruous audio-visual modality. 216 trials (72 items × 3 modalities) were generated for these three modalities. Different tokens were used to create the incongruous audio-visual /r/-/w/ and /th/-/s/ stimuli resulting in 48 video files. The perception experiment thus contained 264 trials.

To reduce experimental time, half of the auditory-only, visual-only and congruous audio-visual trials were presented to a single group of participants (108 trials per group). Each word appeared once in each modality (AO, VO, AVc) across the two groups. Within each group ($n$ = 20), the same word was presented no more than two times. For each phonological contrast under study (/r/-/w/, /l/-/r/, /l/-/w/), the presentation of stimuli was counterbalanced across the two groups. For example, for the /r/-/w/ contrast with the FLEECE vowel presented in the auditory-only modality, while Group 1 heard /r/, Group 2 heard /w/. As previously indicated, the word list contains two sets of minimal pairs for each vowel. The two sets of minimal pairs were also counterbalanced across the two groups. Where one group saw one set, the other group saw the other. So, if we take the same example, the /r/-/w/ contrast with the FLEECE vowel presented in the auditory-only modality, while Group 1 heard *reed*, Group 2 heard *week*. As there were fewer items in the incongruous audio-visual modality than in the other modalities, both groups were presented with all 48 items. There was thus a total of 156 trials per group (108 AO, VO, AVc + 48 AVi). Appendix B.2 presents a full list of the trials generated for the test words for both groups in all four modalities. Further details concerning the creation of perception trials are provided below.

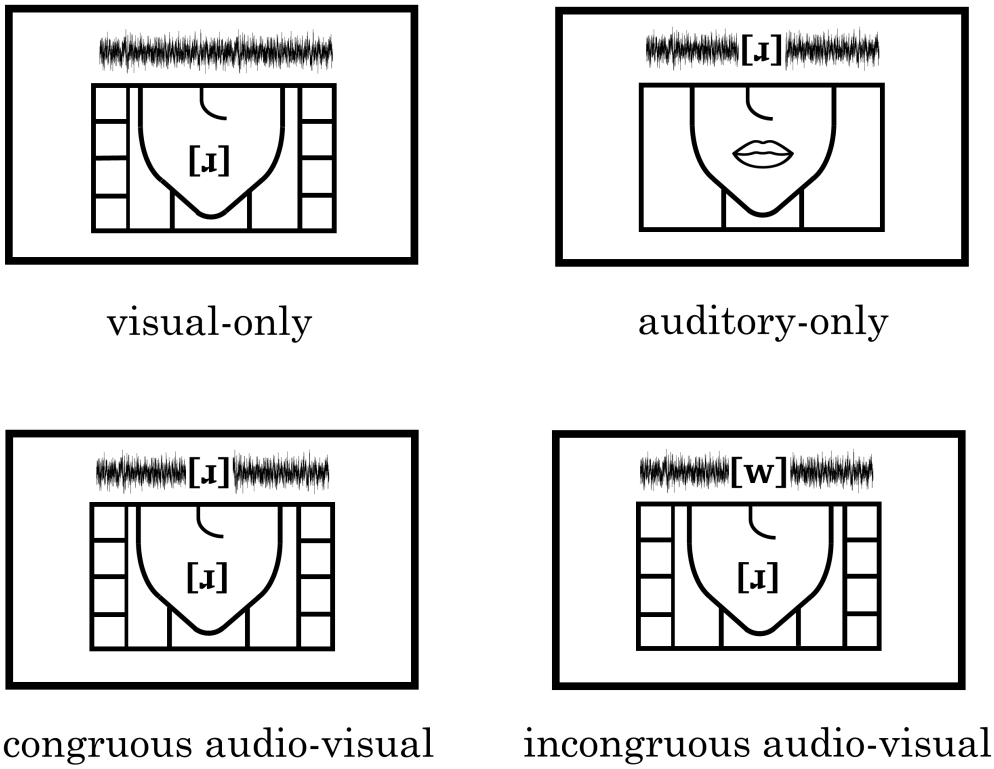



**Figure 6.2:** *A schematic representation of perception trials in all four presentation modalities. The same audio or visual token was never presented more than once. Noise was present in all modalities.*

### Mixing audio with noise

Using VirtualDub, the raw digital video file was cropped to leave only the speaker's face and to position her mouth roughly in the centre of the video pane. The video was resized to a width of 960 and a height of 720 pixels. Individual video clips and their corresponding audio tracks were then extracted and saved separately for each word.

Each audio file was manually segmented at the word level in Praat (Boersma & Weenink, 2019). A Praat script was written to extract the duration of each word and the duration of the intervals of silence surrounding each word. It was agreed that silence intervals should be controlled to match in duration across all perception trials. As a result, the script found the longest silence interval preceding and following all the words in the audio files (926.10 ms and 604.72 ms, respectively). The corresponding values were rounded up to the nearest millisecond. The script then extended the intervals of silence before and after each word to 927 ms and 605 ms, respectively, by calculating the difference between the original lengths of silence and these values. Both the audio files and their corresponding text grids were extended in length.[6]

Auditory stimuli were embedded in pink noise. Pink noise was chosen rather than white noise because it has been found to be the most effective masker of the two (Adachi, Akahane-Yamada, & Ueda, 2006; Rubin-Spitz, McGarr, & Youdelman, 1986). We followed a similar procedure to Havenhill (2018) for the mixing of audio files with pink noise and used the formula provided by Weenink (2014) to generate pink noise in Praat with the following few lines of code:

```
Create Sound from formula: "pinkNoise", 1, 0, 2.4, 44100,
"randomGauss(0 , 1)"
To Spectrum: "no"
Formula: "if x > 100 then self*sqrt(100 / x) else 0 fi"
To Sound
```

The resulting pink noise file was 2.4 seconds long to match the length of the longest audio file

[6]The Praat script used to extract durations is presented in Appendix B.3.

in the dataset. Pink noise was added for the duration of each word file. As we had previously normalised the length of the intervals before and after each word, the onset of noise took place 927 ms before the word began and the offset of noise occurred 605 ms after the end of the word.

Pink noise was mixed with the audio files at a SNR of -12 dB and mean amplitude was scaled to 70 dB using a Praat script adapted from the one by McCloy (2013). Although Havenhill (2018) used an SNR of -15 dB, an SNR of -12 dB was preferable because it has been identified as a 'special zone' where audio-visual benefit is maximal relative to higher and lower SNRs. Ross, Saint-Amour, Leavitt, Javitt, and Foxe (2007) found that audio-visual integration is maximal when the SNR is located between extreme values, where observers have to rely mostly on speech-reading (-24 dB) and where information from visual articulation is largely redundant to the auditory signal (0 dB). Given that the results from the Ross et al. (2007) study suggested that an SNR of -12 dB is the 'sweet spot' (Smayda et al., 2016) for maximum audio-visual multisensory integration, we chose to use the same SNR of -12 dB.

### Auditory-only

For the auditory-only modality, a still image of the speaker's face with a neutral expression (presented in Figure 6.3) was extracted from one of the videos. This still image was then combined with one audio file mixed with pink noise per word. The image of the speaker's face was presented for the duration of each audio file.

**Figure 6.3:** *Still image of the speaker's face with a neutral expression presented during the auditory-only modality.*

### Visual-only

Video files were extended to match the duration of their corresponding audio files, so that the intervals of silence preceding and following each word were a constant length (927 ms and 605 ms, respectively). The video files were extended using a script written in Matlab. To extend the length of the interval occurring before the word, the script selected the very first frame of each video and repeated it for the necessary length of time. To extend the interval after the word, the last video frame was extended in the same manner. Visual-only trials were created from the same tokens as the auditory-only trials. For visual-only, each extended video file was dubbed with the plain pink noise audio file generated in Praat (as described in Section 6.2.3) with mean amplitude scaled to 70 dB. The video and the pink noise started and stopped simultaneously. As a result, noise onset was at the same time as video onset, 927 ms before the word began. Noise offset occurred 605 ms after the end of the word, coinciding with the end of the video.

### Congruous audio-visual

To create congruous audio-visual trials, a different token to the one used for the auditory-only and visual-only trials was used for each word. Extended video files were combined with their

corresponding extended audio files mixed with pink noise. As with the visual-only modality, noise onset and video onset occurred simultaneously, 927 ms before the word began. Noise offset occurred 605 ms after the end of the word, coinciding with the end of the video.

#### Incongruous audio-visual

Incongruous trials containing auditory /r/ paired with visual /w/, and auditory /w/ paired with visual /r/ were produced using Virtual Dub. /s/-/th/ incongruous audio-visual pairings were also created to act as controls. 24 trials were created from the word-initial /r/ and /w/ minimal pairs presented in Table 6.2 (12 × incongruous audio-visual /r/-/w/ + 12 × incongruous audio-visual /w/-/r/). The same number of /s/-/th/ and /th/-/s/ incongruous audio-visual trials were produced from the filler and control words (presented in Appendix B.1). Audio files in which the length of pre- and post-word silence intervals matched in length, were mixed with pink noise. Different tokens to the ones presented in the other modalities were used to create incongruous audio-visual trials. Incongruous audio-visual word pairings were matched as closely as possible in word length (mean difference = 8.65 ± 6.41 ms). As all the words began at the same time (927 ms from the onset of the recording) and the words were matched closely in length, the dubbing procedure for incongruous audio-visual trials was identical to that of congruous ones: the audio and video were set to occur simultaneously.

### *6.2.4 Procedure*

The perception experiment took place in a quiet room either at the University of York or in the participant's home or workplace. Participants were seated in front of a portable laptop computer with a 13 inch screen. Audio was presented through a pair of AKG K271 headphones with the volume set to a comfortable level. The experiment was carried out using PsychoPy (Peirce, 2007). Participants were informed that the purpose of the study was to investigate how English listeners perceive and decode speech when the audio is masked with noise. They did not know that the main study interest involved English /r/, nor were they informed that they would be presented with incongruous audio-visual trials.

Following the methodology described in Ross et al. (2007), stimulus presentation of auditory-only, visual-only, congruous and incongruous audio-visual trials were randomly intermixed. Trial order was unique to each participant. An image or video of the speaker's face was displayed for the total length of each audio file in the middle of the screen, preceded by a fixation cross of duration 2 000 ms. The fixation cross was placed to coincide with the level of the speaker's lips in the stimuli and participants were instructed to look directly at the cross while it was on the screen. Directly after stimulus presentation, participants identified the word-initial consonant they perceived by selecting a word from two options presented on screen. The two words appeared in alphabetical order (i.e., words beginning with 'r' and 'w', 'l' and 'r', 'l' and 'w') and were separated by a large space. The two word options were positioned to align vertically with the level of the speaker's lips. Following Havenhill (2018), a 2 000 ms time limit was imposed on responses, after which the programme automatically advanced to the next stimulus. Participants selected their response by clicking on the word using a wireless optical mouse. They were instructed that their first mouse click would be recorded and were asked to respond as quickly and accurately as possible.[7] We chose to use a mouse rather than key presses because we felt that participants would be less likely to avert their gaze from the screen when the mouse cursor was visible. Participants were given 10 practice trials in which minimal pair /hVd/ words produced by the same speaker were presented in the four modalities. Participants were provided with four equally-distributed, self-timed breaks. The experiment took around 20 minutes to complete. Figure 6.4 presents a schematisation of the perception experiment design.

[7]The exact instructions we gave to perception participants are presented in Appendix B.6.

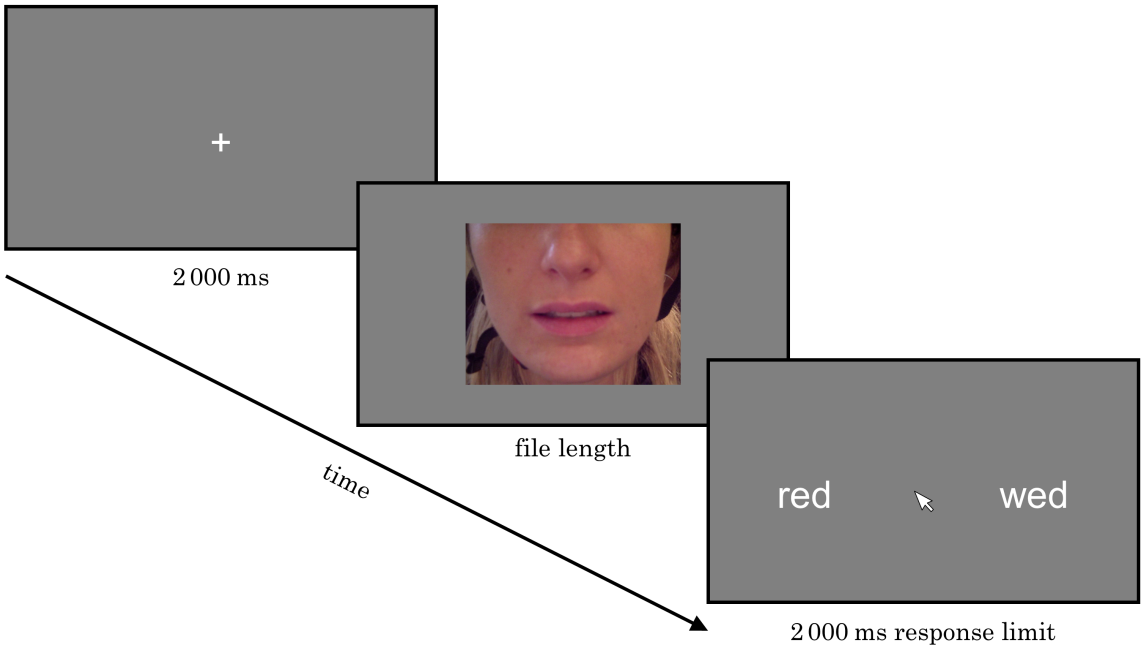



**Figure 6.4:** *Perception experiment design. The longest file was 2 400 ms long.*

### Catch trials

In addition to the experimental conditions, 10 catch trials were presented to enforce ongoing attention to the video (as described in Irwin, Frost, Mencl, Chen, & Fowler, 2011). An example catch trial was also included in the practice round before the experiment began. Catch trials consisted of a random auditory stimulus from the dataset[8] mixed with pink noise accompanied by a still image of the speaker in which her lips were painted in a bright colour. Participants were instructed to respond with the colour of her lips and not the word she said in these cases. As a result, the two possible responses to a catch trial were the colour of her lips or the word she said. An example image from a catch trial is presented in Figure 6.5, for which one of the responses was 'orange'.

[8]The auditory stimuli for catch trials came from tokens which had not already been used in the experimental conditions.

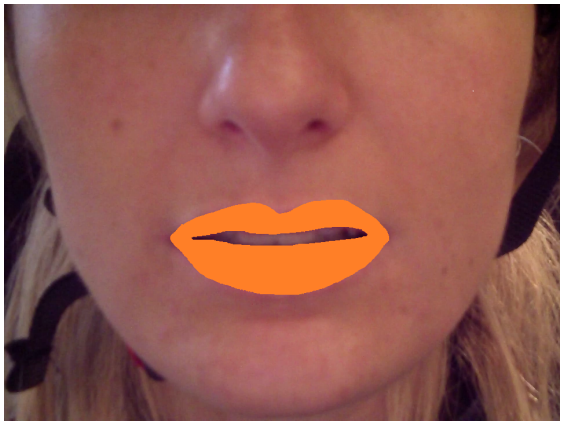


**Figure 6.5:** *An image from a catch trial in which the speaker's lips were painted in a colour. Participants were instructed to respond with the colour of the speaker's lips and not the word they perceive.*

### 6.2.5 Statistical analysis

As in the previous two experiments, statistical analysis of perception data was implemented in R using the lmer() function of the *lme4* package to perform a series of linear mixed-effects models. We tested the significance of main effects to model fit using likelihood ratio tests. Model fit was assessed with a comparison of Akaike Information Criterion (AIC). Model residuals were plotted to test for deviations from homoscedasticity or normality.

### 6.2.6 Analysis of production data

In order to verify that the Anglo-English speaker who produced stimuli for the perception experiment converged with the speakers we recorded in our production experiments, the labial articulation and acoustics of her /r/ and /w/ productions were analysed. We chose not to run statistical analysis on the results from these analyses because it was felt that statistical power would not be sufficient.

### Acoustic analysis

Praat's Burg algorithm was used to obtain formant values for /r/ and /w/. The ceiling of the formant search range in Hertz was set to 5 500 Hz, which is the recommended limit for an adult female voice. Using Praat's formant detection, the minimum F3 value in each /r/ token and the minimum F2 value in each /w/ token were labelled. We verified that formant estimation matched the underlying spectrogram at these points and the first three formants (F1-F3) were extracted. 60 tokens of word-initial /r/ and 60 tokens of word-initial /w/ were analysed in the following vowel contexts: FLEECE, KIT, DRESS, TRAP, FACE, PRICE.

### Lip measures

In order to measure the speaker's lips in world units, a physical ruler was placed directly below the speaker's lips touching her chin, which was recorded before she read out the word list. One video frame presenting an image of the ruler was extracted from the video and opened in ImageJ (version 1.52) (Schneider, Rasband, & Eliceiri, 2012). A straight line was positioned from 0 to 10 cm along the image of the physical ruler, which produced a distance of 816 pixels. This distance was used to set a global measurement scale of 81.6 pixels/cm for all subsequent measures carried out in ImageJ. An image depicting the straight line positioned along the ruler in ImageJ is presented in Figure 6.6. Video files for each token were then opened in ImageJ and the image presenting maximum labial constriction was selected by holistically examining sequential video frames. Lip width was measured by placing a quasi-horizontal line from lip corner to corner. Lip aperture was measured by positioning a straight vertical line from the vermilion border of the top lip just below the philtrum dimple down to the vermilion border of the bottom lip. Example images of lip width and aperture measurements are presented in Figure 6.7, which depicts an /r/ token. Using ImageJ, the length and position of these two lines were measured in centimetres. The position of the lip aperture (vertical) line along the y-axis was used to measure the vertical position of the lips.[9] As with the previous manual lip

[9]The lip aperture line was chosen over the lip width line as the lip aperture line was always perfectly straight, contrary to the lip width line.

**Figure 6.6:** *Lip dimension measures in world units. A straight horizontal line (in yellow) was positioned along an image of a physical ruler for a distance of 10 cm.*

**Figure 6.7:** *Lip width (left) and lip aperture (right) lines (in yellow) positioned for lip dimension measures.*

measures presented in this thesis, the lips were also measured in a neutral setting.[10] Measures for /r/ and /w/ could then be compared with the neutral lip setting.

[10]The neutral lip image we presented in the visual-only modality in the perception experiment (presented in Figure 6.3) was used to measure the neutral lip setting.

## 6.3 RESULTS

### *6.3.1 Production data*

#### Acoustics of /r/ and /w/

Table 6.3 presents the average first, second and third formant values for /r/ and /w/ (in Hz) along with their standard deviations from the auditory data collected from the native Anglo-English speaker who provided stimuli for the perception experiment. Formant values for this speaker are presented alongside those measured in the previous study (Experiment 2), in which the data from 23 speakers' productions of /r/ and /w/ were analysed.

The formant values acquired across the two experiments are very similar. We notice that the average third formant for /r/ is somewhat higher in the perception speaker than the average F3 from Experiment 2. This may be due to the fact that the perception speaker did not produce any tokens of /r/ in the context of back vowels, unlike the production experiment speakers. We know from Experiment 1 that F3 is at its lowest in the context of open-back vowels as opposed to close-front ones. However, the perception speaker's average formant values for /r/ still lie within the normal range reported in previous studies on English /r/ (300-500 Hz for F1, 900-1 300 Hz for F2, and 1 300-2 000 Hz for F3, as discussed in Chapter 2, Section 2.8, p. 64). As with the speakers in Experiment 2, these results confirm that both F2 and F3 differ substantially for /r/ and /w/ in the speaker presented in the perception experiment. While the speaker's /w/ productions produce an average F2 that is 380 Hz lower than that of /r/, F3 is 628 Hz higher for /w/ than /r/.

| Phoneme | Experiment | F1 | F2 | F3 |
|---|---|---|---|---|
| /r/ | 2 | 418 (65) | 1 242 (226) | 1 900 (212) |
| | 3 | 439 (21) | 1 169 (109) | 2 004 (102) |
| /w/ | 2 | 401 (66) | 743 (171) | 2 716 (241) |
| | 3 | 407 (66) | 789 (150) | 2 632 (238) |

**Table 6.3:** *Mean formant values (in Hz) and their standard deviations (in parentheses) for* /r/ *and* /w/ *produced by the speaker who supplied stimuli for the present perception experiment (Experiment 3) and by the 23 speakers from the production study (Experiment 2).*

### Lip measures

Table 6.4 presents descriptive statistics for lip width, height and vertical position in mm for /r/ and /w/ in the speaker who provided perception stimuli. The values observed in a neutral lip setting are also presented. As in Experiment 2, lower values in the vertical lip position measure correspond to a *higher* lip position. Based on the lip measures obtained automatically using a Convolutional Neural Network in Experiment 2, the strongest predictors of phoneme category were lip width and vertical position. It was observed that the lips are significantly wider and significantly higher for /r/ than they are for /w/. We observe the same pattern in the perception speaker. On average, the lips are over half a centimetre (6.58 mm) wider and 4 mm higher for /r/ than they are for /w/. While lip height increases for both /r/ and /w/ from the neutral lip setting, there is little difference between their average values. We also observe that while /r/ has a higher vertical lip position on average than the neutral setting, vertical lip position lowers for /w/. Like the speakers from Experiment 2, these results confirm that the labial gestures for /r/ and /w/ differ in the perception speaker, predominantly with regards to lip width and lip position; /r/ has a wider and higher lip posture than /w/. We will now assess whether or not these differences are perceptible to native speakers of Anglo-English.

| Phoneme | Width | Height | Vertical position |
|---|---|---|---|
| neutral | 44.30 | 13.90 | 55.60 |
| /r/ | 45.84 (1.08) | 14.72 (0.81) | 53.43 (0.98) |
| /w/ | 39.26 (1.57) | 15.23 (0.70) | 57.43 (1.16) |

**Table 6.4:** *Mean lip dimensions (in mm) and their standard deviations (in parentheses) for* /r/, /w/ *and a neutral lip setting in the speaker who supplied stimuli for the perception experiment. Lower values in vertical position correspond to a higher lip position.*

### 6.3.2 *Responses to catch trials*

As previously stated in Section 6.2.4 (p. 230), 10 catch trials were included in the perception experiment to enforce ongoing attention to the video and to confirm that participants attended to visual cues. Figure 6.8 presents the number of correct responses to catch trials per participant. All subjects but one correctly responded to at least 7/10 catch trials. Subject 21 accurately responded to only two catch trials. This subject reported to have an uncorrected sight problem. As a result, we decided to exclude his responses from subsequent analyses.

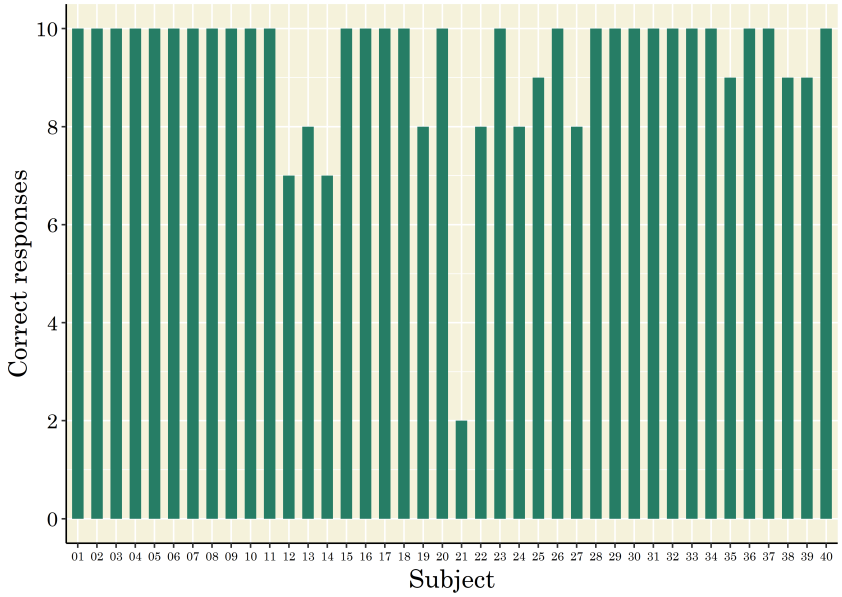



**Figure 6.8:** *Number of correct responses per participant to 10 catch trials.*

### 6.3.3 *Perception of unimodal and congruous audio-visual trials*

Participants responded to 54 unimodal (auditory-only, visual-only) and congruous audio-visual trials presenting monosyllabic words beginning with /r/, /w/ and /l/ in noise, resulting in a total of 2106 observations (39 subjects × 54 trials). Table 6.5 presents stimulus-response confusion matrices for the three modalities. As the table indicates, some trials ($n$ = 126) were left unanswered, particularly in the visual-only condition. A closer inspection of the no response trials indicated that they came from 26 out of the 39 subjects. Among these 26 subjects, the average number of unanswered trials was 4.85 ± 3.72. The highest number of unanswered trials in any participant was 19/54. The 126 unanswered trials were excluded from statistical analysis resulting in 1 980 analysable observations.

| | Presented | | | | | |
|---|---|---|---|---|---|---|
| | auditory-only | | visual-only | | audio visual | |
| | /l/ | /w/ | /l/ | /w/ | /l/ | /w/ |
| Responded 'l' | 91 | 44 | 79 | 4 | 112 | 1 |
| Responded 'w' | 15 | 68 | 16 | 107 | 3 | 114 |
| No response | 11 | 5 | 22 | 6 | 2 | 2 |
| | /l/ | /r/ | /l/ | /r/ | /l/ | /r/ |
| Responded 'l' | 100 | 37 | 75 | 4 | 115 | 1 |
| Responded 'r' | 14 | 77 | 22 | 102 | 1 | 116 |
| No response | 3 | 3 | 20 | 11 | 1 | 0 |
| | /r/ | /w/ | /r/ | /w/ | /r/ | /w/ |
| Responded 'r' | 100 | 55 | 103 | 6 | 112 | 11 |
| Responded 'w' | 12 | 51 | 3 | 107 | 1 | 101 |
| No response | 5 | 11 | 11 | 4 | 4 | 5 |

**Table 6.5:** *Raw stimulus-response confusion matrices for the identification of* /r/, /w/ *and* /l/ *in unimodal and congruous audio-visual modalities.*

As described in Section 6.1.3 (p. 215), a common technique for quantifying the benefit obtained from adding a visual signal to an auditory stimulus (i.e., visual enhancement) is to calculate the difference between a listener's performance in audio-visual and auditory-only conditions expressed relative to the amount of possible improvement given the subject's

auditory-only score. While we had hoped that the addition of pink noise would make the implementation of this measure possible by allowing participants room for improvement, many of the participants obtained perfect accuracy scores in the auditory-only condition, despite the addition of noise, which made it impossible to run.

As we were unable to measure visual enhancement, accuracy scores were instead converted to the sensitivity measure $d'$ (d-prime) from Signal Detection Theory (Green & Swets, 1966). These measures were implemented in a similar manner to those presented in McGuire and Babel (2012), in which the contribution of audio-visual cues to the sound change involving the English /f/-/θ/ contrast was considered using a similar design to the one employed in the present study. Sensitivity measures allow for a more accurate comparison across conditions and subjects than comparisons of the proportion of correct responses (McGuire & Babel, 2012). Signal Detection Theory is often associated with classification experiments in which participants judge whether a stimulus is present or absent, responding with 'yes' or 'no' (i.e., 'yes, the stimulus was present' or 'no, the stimulus was not present'). In phonetic studies, this sort of classification experiment tends to occur in AX discrimination tasks, in which subjects are asked to judge whether a pair of stimuli are the 'same' or 'different'. To measure sensitivity, Signal Detection Theory considers the probability that a subject says 'yes' when a stimulus is present (*hit rate*) but also the probability that the subject says 'yes' when the stimulus is absent (*false alarm rate*). However, Signal Detection Theory may be applied to any perceptual experiment in which two different types of stimuli must be discriminated (Stanislaw & Todorov, 1999). As a result, following McGuire and Babel (2012), the hit and false alarm rates were calculated for each of the three contrasts /l/-/w/, /l/-/r/, /r/-/w/ in each of the three modalities per subject by arbitrarily assigning correct responses for one of the phonemes in each pair as hits. Hits were assigned to correct /l/ responses in the /l/-/w/ and the /l/-/r/ contrasts, and to correct /r/ responses in the /r/-/w/ contrast. False alarms were assigned to incorrect responses of the same phonemes in each of the respective contrasts. An example of the categorisation of trials and responses into hits, misses, false alarms and correct rejections is presented in Table 6.6 for the /r/-/w/ contrast. The following equations were used to calculate hit and false alarm rates

and $d'$ for each contrast in each modality per subject (Macmillan & Creelman, 2005):

$$\text{hit rate } H = \frac{\text{hits}}{(\text{hits} + \text{misses})}$$

$$\text{false alarm rate } F = \frac{\text{false alarms}}{(\text{false alarms} + \text{correct rejections})}$$

$$d' = z(H) - z(F)$$

For the contexts in which a subject attained perfect accuracy, hit and false alarm rates of 0 and 1 were converted to $1/(2N)$ and $1 - 1/(2N)$ respectively, where $N$ is the number of trials on which the proportion is based (Macmillan & Creelman, 2005). A $d' = 0$ indicates that a subject has no sensitivity to a contrast, i.e., that the subject is responding randomly (McGuire & Babel, 2012). The maximal $d'$ score in the dataset was just over 2.9, which we consider to be near perfect perception. The percentage of correct responses was also calculated by dividing the number of correct responses (hits + correct rejections) by the total number of responses (hits + misses + false alarms + correct rejections).

| | Stimulus: /r/ | Stimulus: /w/ |
|---|---|---|
| Response: 'r' | hit | false alarm |
| Response: 'w' | miss | correct rejection |

**Table 6.6:** *Categorisation of hits, misses, false alarms and correct rejections in the* /r/-/w/ *and* /w/-/r/ *stimulus-response pairs.*

An initial inspection of the stimulus-response matrices presented in Table 6.5 reveals that in the auditory-only trials, while subjects were generally able to accurately identify /r/ and /l/ tokens, the proportion of correctly identified /w/ tokens was comparatively lower. In actual fact, when presented with /w/ audio stimuli in the context of /r/ distractors, subjects selected more 'r' than 'w' responses overall, which suggests there may be a preference for /r/ in this particular modality. However, this apparent response bias does not appear to extend to the

other modalities for the /r/-/w/ contrast. To measure response bias, Criterion Location ($c$) was calculated using the following formula (Macmillan & Creelman, 2005):

$$\text{response bias } c = -\frac{1}{2}[z(H) + z(F)]$$

Table 6.7 reports summary statistics for the three contrasts /l/-/w/, /l/-/r/, /r/-/w/ in the three modalities. The bias measure ($c$) indeed reflects our observation from the raw data that 'r' responses were more likely than 'w' responses for the /r/-/w/ contrast in the auditory-only (AO) modality because it results in the mean Criterion Location value which is furthest from zero out of all contrasts and modalities (-0.53). As predicted, the highest sensitivity to contrasts, as well as the highest proportions of correct responses, occurred in the audio-visual (AVc) modality. Interestingly, the sensitivity and proportion correct measures indicate very little difference in perception between the visual-only and the audio-visual modalities for the /r/-/w/ contrast. Figure 6.9 presents plots of the proportion of correct and incorrect responses for each of the three contrasts in each of the three modalities, which paints the same picture.

| Contrast | Modality | Sensitivity | Bias | Proportion correct |
|---|---|---|---|---|
| /l/-/w/ | AO | 1.26 (0.62) | -0.33 (0.43) | 0.73 (0.12) |
| | VO | 2.02 (0.87) | 0.19 (0.32) | 0.89 (0.15) |
| | AVc | 2.71 (0.39) | 0.04 (0.17) | 0.98 (0.06) |
| /l/-/r/ | AO | 1.51 (0.86) | -0.27 (0.42) | 0.78 (0.16) |
| | VO | 1.90 (1.02) | 0.23 (0.43) | 0.87 (0.19) |
| | AVc | 2.60 (0.27) | 0.00 (0.13) | 0.99 (0.04) |
| /r/-/w/ | AO | 1.05 (1.01) | -0.53 (0.47) | 0.69 (0.19) |
| | VO | 2.44 (0.54) | -0.04 (0.27) | 0.96 (0.09) |
| | AVc | 2.43 (0.03) | -0.12 (0.32) | 0.95 (0.12) |

**Table 6.7:** *Summary statistics for each contrast in each modality presenting mean and standard deviation (in parentheses) values for each measure: sensitivity in d′, response bias in c (0 = no bias, negative indicates bias to respond with the first phoneme presented in each contrast pair), and the proportion of correct responses.*

To test whether there are statistically significant differences in the perception of /r/, /w/

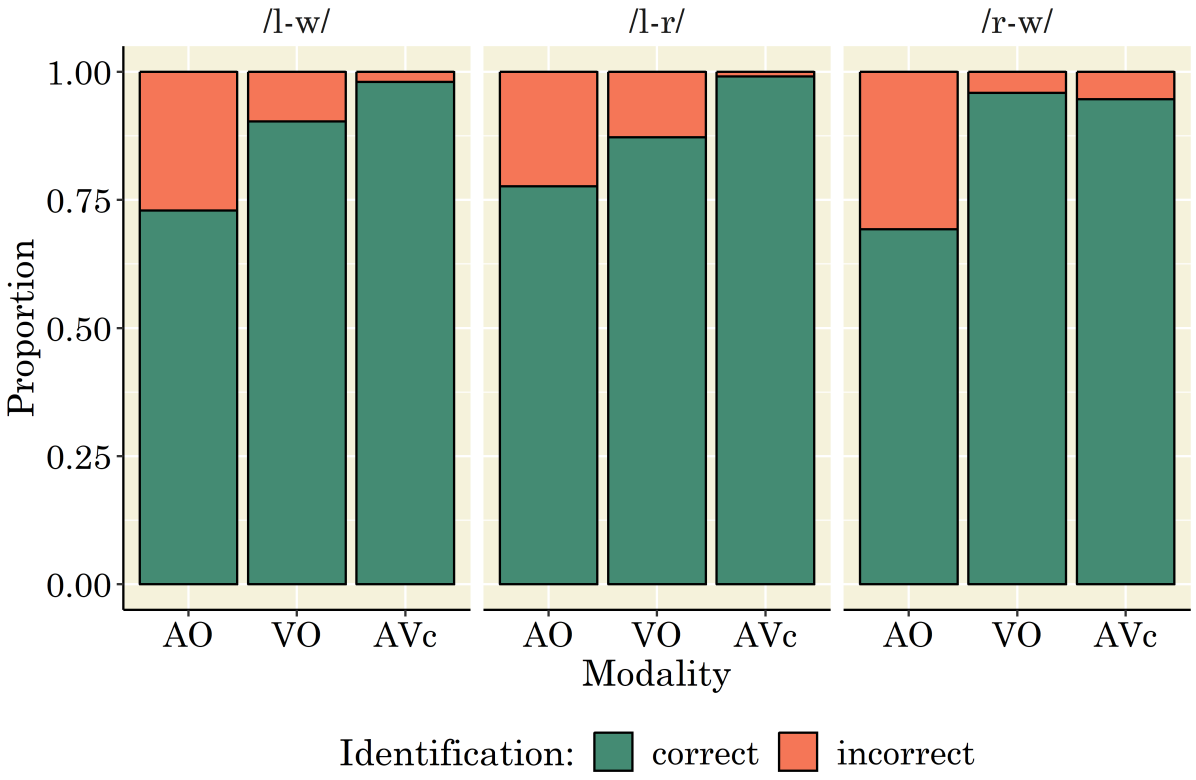



**Figure 6.9:** *Proportion of correct and incorrect responses for* /l/-/w/, /l/-/r/, /r/-/w/ *contrasts in unimodal and congruous modalities.*

and /l/ in the three modalities, we performed a series of linear mixed-effects analyses. It was predicted that if /r/ has a perceptible labial gesture, the perception of /r/ stimuli should not significantly differ from that of /w/ presented in the visual-only condition. On the other hand, if /r/ does not have a perceptible labial component, /r/ should not significantly differ from /l/. We therefore implemented a generalised linear mixed-effects model predicting the correct perception of /l/, /w/ and /r/ stimuli presented in auditory-only, visual-only and audio-visual modalities. Accuracy was the binary outcome variable (incorrect versus correct) that was regressed against stimulus (/l/, /w/, /r/) and modality (AO, VO, AVc) with an interaction term. Other fixed factors also included subject sex (female, male), age, origin (England or abroad) and their hearing score (as presented in the participant demographics in Table 6.1, p. 218). Sex was included as a factor because previous studies have observed differences in perception between men and women, as discussed in Chapter 1. The numeric fixed factors of age and hearing score were converted to z-scores. The maximal set of successfully converging random slopes and intercepts for subjects and items were included, which turned out to be random intercepts for subjects and items. The addition of random slopes failed to converge.

Likelihood ratio tests revealed that the interaction between stimulus and modality was highly significant ($\chi^2(4) = 89.37$, $p < .001$). The main effects of both stimulus and modality were also significant (Stimulus: $\chi^2(2) = 6.92$, $p = 0.03$, Modality: $\chi^2(2) = 170.61$, $p < .001$). Subject sex and origin were also significant main effects (Sex: $\chi^2(1) = 5.46$, $p = 0.02$, Origin: $\chi^2(1) = 5.23$, $p = 0.02$). However, subject age and hearing score failed to reach significance (Age: $\chi^2(1) = 2.34$, $p = 0.13$, Hearing: $\chi^2(1) = 0.62$, $p = 0.43$). For the significant interaction between stimulus and modality, pairwise Tukey post-hoc tests were performed on all possible combinations using the *lsmeans* package (Lenth, 2016). The model summary for the best fitting final model is presented in Table 6.8. Table 6.9 reports pertinent pairwise comparisons of the interaction between stimulus and modality. Figure 6.10 presents a plot of the predicted probability of accurately identifying each stimulus across the three modalities according to the best-fitting model.

The results from this analysis indicate that identification of /w/ and /r/ tokens significantly

| Predictor | Estimate (log-odds) | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 1.70 | 0.44 | 3.86 | $< .001^{***}$ |
| Stimulus /w/ | -1.97 | 0.45 | -4.43 | $< .001^{***}$ |
| Stimulus /r/ | -0.63 | 0.46 | -1.38 | 0.17 |
| Modality VO | -0.65 | 0.29 | -2.27 | $0.02^{*}$ |
| Modality AVc | 1.99 | 0.49 | 4.03 | $< .001^{***}$ |
| Sex male | -0.52 | 0.19 | -2.73 | $0.01^{**}$ |
| Origin England | 0.78 | 0.33 | 2.36 | $0.02^{*}$ |
| Stimulus /w/: Modality VO | 3.86 | 0.47 | 8.27 | $< .001^{***}$ |
| Stimulus /r/: Modality VO | 2.95 | 0.51 | 5.75 | $< .001^{***}$ |
| Stimulus /w/: Modality AVc | 1.00 | 0.60 | 1.67 | 0.10 |
| Stimulus /r/: Modality AVc | 1.67 | 0.88 | 1.91 | 0.06 |

$$Accuracy \sim Stimulus \times Modality + Sex + Origin + (1|Subject) + (1|Item)$$

**Table 6.8:** *Output of a generalised linear mixed-effects model predicting the probability a token is accurately identified. The intercept represents /l/ stimuli in the AO modality perceived by a female subject who grew up abroad.*

| Contrast | | Odds ratio | Std. Error | *t* value | *p* value |
|---|---|---|---|---|---|
| /l/ | AO–VO | 1.91 | 0.54 | 2.27 | 0.36 |
| | AO–AVc | 0.14 | 0.07 | -4.03 | $0.002^{**}$ |
| | VO–AVc | 0.07 | 0.04 | -5.37 | $< .001^{***}$ |
| /w/ | AO–VO | 0.04 | 0.01 | -8.73 | $< .001^{***}$ |
| | AO–AVc | 0.05 | 0.02 | -8.77 | $< .001^{***}$ |
| | VO–AVc | 1.24 | 0.55 | 0.49 | 0.99 |
| /r/ | AO–VO | 0.10 | 0.04 | -5.47 | $< .001^{***}$ |
| | AO–AVc | 0.03 | 0.02 | -5.07 | $< .001^{***}$ |
| | VO–AVc | 0.26 | 0.20 | -1.70 | 0.74 |
| AO | /l/–/w/ | 7.16 | 3.19 | 4.42 | $< .001^{***}$ |
| | /l/–/r/ | 1.87 | 0.85 | 1.38 | 0.91 |
| | /r/–/w/ | 0.26 | 0.11 | 3.11 | $0.048^{*}$ |
| VO | /l/–/w/ | 0.15 | 0.08 | -3.59 | $0.010^{*}$ |
| | /l/–/r/ | 0.10 | 0.06 | -4.07 | $0.002^{**}$ |
| | /r/–/w/ | 0.64 | 0.41 | -0.70 | 0.99 |
| AVc | /l/–/w/ | 2.63 | 1.72 | 1.50 | 0.87 |
| | /l/–/r/ | 0.35 | 0.32 | -1.14 | 0.97 |
| | /r/–/w/ | 0.13 | 0.11 | 2.37 | 0.30 |

**Table 6.9:** *Post-hoc pairwise comparisons of the significant interaction between Stimulus and Modality on identification accuracy from a generalised linear mixed-effects model.*

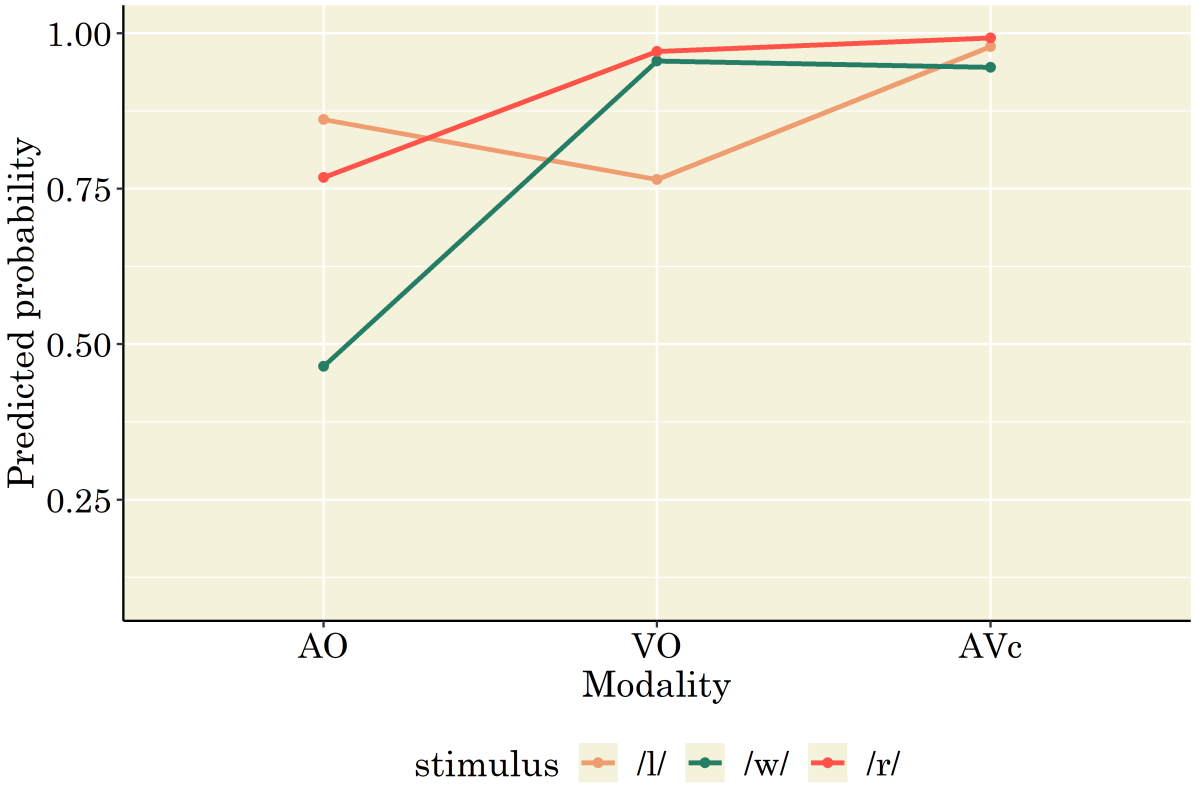



**Figure 6.10:** *Predicted probability of correctly identifying* /l/, /w/ *and* /r/ *stimuli in each modality from a generalised linear mixed-effects model.*

improves when subjects are presented with the visual cue of the bottom half of the speaker's face. The probability of accurately identifying /w/ and /r/ tokens is significantly higher in the visual-only and the audio-visual modalities than the auditory-only one. Previous studies have generally observed that participants are more successful at identifying speech in auditory-only than in visual-only conditions. However, the results from the perception of /r/ and /w/ stimuli not follow this trend. The model predicts perceptual accuracy to be nearly perfect in the visual-only modality for /w/ (0.96 ± 0.02) and for /r/ (0.97 ± 0.01). In contrast, accuracy is significantly lower in the auditory-only modality than in the visual-only one for /w/ and /r/ and accuracy is actually predicted to be lower than chance for /w/ (0.46 ± 0.08). However, the model predicts accuracy to be well above chance for both /l/ (0.86 ± 0.04) and /r/ (0.77 ± 0.06), which allows us to conclude that despite the addition of pink noise, participants were still sensitive to acoustic cues. Indeed, while the presence of the visual cue aided the identification of /w/ and /r/, /l/ tokens were best identified when the auditory cue was present. As Figure 6.10 indicates, the probability of accurately identifying /l/ stimuli is lower in the visual-only (0.77 ± 0.06) than in the auditory-only (0.86 ± 0.04) modality, although this difference did not reach significance. Contrary to /r/ and /w/, the probability of accurately identifying /l/ tokens significantly improves from the visual-only modality with the presence of auditory cues in the audio-visual one. /l/ is thus the only phoneme to benefit from the combination of both auditory and visual cues. These results therefore indicate that the perception of /r/ patterns with that of /w/ across all three modalities. This is important because it suggests that /r/, like /w/, has a perceptible labial gesture.

Although accuracy was particularly high for both /w/ and /r/ stimuli in the visual-only modality, we do not yet know to what extent subjects were able to distinguish between the labial configurations of /r/ and /w/. As a result, another linear mixed-effects analysis was implemented predicting subjects' sensitivity to each of the three contrasts in the three presentation modalities. In this model, $d'$ was the outcome variable which was regressed against contrast (/l/-/r/, /l/-/w/, /r-w) and modality (AO, VO, AVc) with an interaction term. Like in the previous model, subject sex, age, origin and hearing were also included as fixed effects. The

numeric fixed effects were again converted to z-scores. Random intercepts for subjects were included.

Likelihood ratio tests revealed that the interaction between contrast and modality was highly significant ($\chi^2(4) = 23.20$, $p = 0.001$). The main effect of modality was also highly significant ($\chi^2(2) = 161.2$, $p < .001$), while the main effect of contrast failed to reach significance ($\chi^2(2) = 0.15$, $p < 0.93$). The main effects of subject sex and origin were also significant (sex: $\chi^2(1) = 3.87$, $p < 0.05$, origin: $\chi^2(1) = 7.11$, $p = 0.008$). Neither hearing score nor age reached significance (hearing: $\chi^2(1) = 0.38$, $p = 0.54$, age: $\chi^2(1) = 2.86$, $p = 0.09$). For the significant interaction between contrast and modality, pairwise Tukey post-hoc tests were again performed on all possible combinations and Table 6.11 presents the most pertinent comparisons for this study. The model summary for the best fitting model is presented in Table 6.10 and Figure 6.11 presents plots of predicted sensitivity to each contrast in the three modalities according to the model.

| Predictor | Estimate | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 0.94 | 0.19 | 4.93 | $< .001^{***}$ |
| Contrast /l/-/r/ | 0.25 | 0.16 | 1.56 | 0.12 |
| Contrast /r/-/w/ | -0.21 | 0.16 | -1.33 | 0.19 |
| Modality VO | 0.76 | 0.16 | 4.71 | $< .001^{***}$ |
| Modality AVc | 1.45 | 0.16 | 8.99 | $< .001^{***}$ |
| Sex male | -0.21 | 0.09 | -2.28 | $0.03^{*}$ |
| Origin England | 0.45 | 0.17 | 2.63 | $0.012^{*}$ |
| Contrast /l/-/r/: Modality VO | -0.37 | 0.23 | -1.61 | 0.11 |
| Contrast /r/-/w/: Modality VO | 0.63 | 0.23 | 2.77 | 0.006** |
| Contrast /l/-/r/: Modality AVc | -0.36 | 0.23 | -1.58 | 0.11 |
| Contrast /r/-/w/: Modality AVc | -0.07 | 0.23 | -0.31 | 0.76 |

*Sensitivity* ~ *Contrast* × *Modality* + *Sex* + *Origin* + (1|*Subject*)

**Table 6.10:** *Output of a linear mixed-effects model predicting perceptual sensitivity (d′). The intercept represents stimuli in the* /l/-/w/ *contrast presented in the AO modality perceived by a female subject who grew up abroad.*

Despite the auditory cues being masked in noise, our statistical analysis indicates that participants were sensitive to acoustic cues in all three contrasts because the regression model

| Contrast | | Estimate | $t$ value | $p$ value |
|---|---|---|---|---|
| /l/-/r/ | AO–VO | -0.39 | -2.42 | 0.27 |
| | AO–AVc | -1.09 | -6.75 | $< .001^{***}$ |
| | VO–AVc | -0.70 | -4.33 | $< .001^{***}$ |
| /l/-/w/ | AO–VO | -0.79 | -4.71 | $< .001^{***}$ |
| | AO–AVc | -1.45 | -9.00 | $< .001^{***}$ |
| | VO–AVc | -0.69 | -4.28 | $< .001^{***}$ |
| /r/-/w/ | AO–VO | -1.39 | -8.662 | $< .001^{***}$ |
| | AO–AVc | -1.38 | -8.56 | $< .001^{***}$ |
| | VO–AVc | 0.01 | 0.07 | 0.99 |
| AO | /l-r/–/l-w/ | 0.25 | 1.56 | 0.83 |
| | /l-r/–/r-w/ | 0.46 | 2.88 | 0.10 |
| | /l-w/–/r-w/ | 0.21 | 1.33 | 0.92 |
| VO | /l-r/–/l-w/ | -0.12 | -0.73 | 0.99 |
| | /l-r/–/r-w/ | -0.53 | -3.32 | $0.03^{*}$ |
| | /l-w/–/r-w/ | -0.42 | -2.59 | 0.20 |
| AVc | /l-r/–/l-w/ | -0.11 | -0.68 | 0.99 |
| | /l-r/–/r-w/ | 0.17 | 1.08 | 0.98 |
| | /l-w/–/r-w/ | 0.28 | 1.76 | 0.71 |

**Table 6.11:** *Post-hoc pairwise comparisons of the significant interaction between Contrast and Modality on perceptual sensitivity from a linear mixed-effects model.*

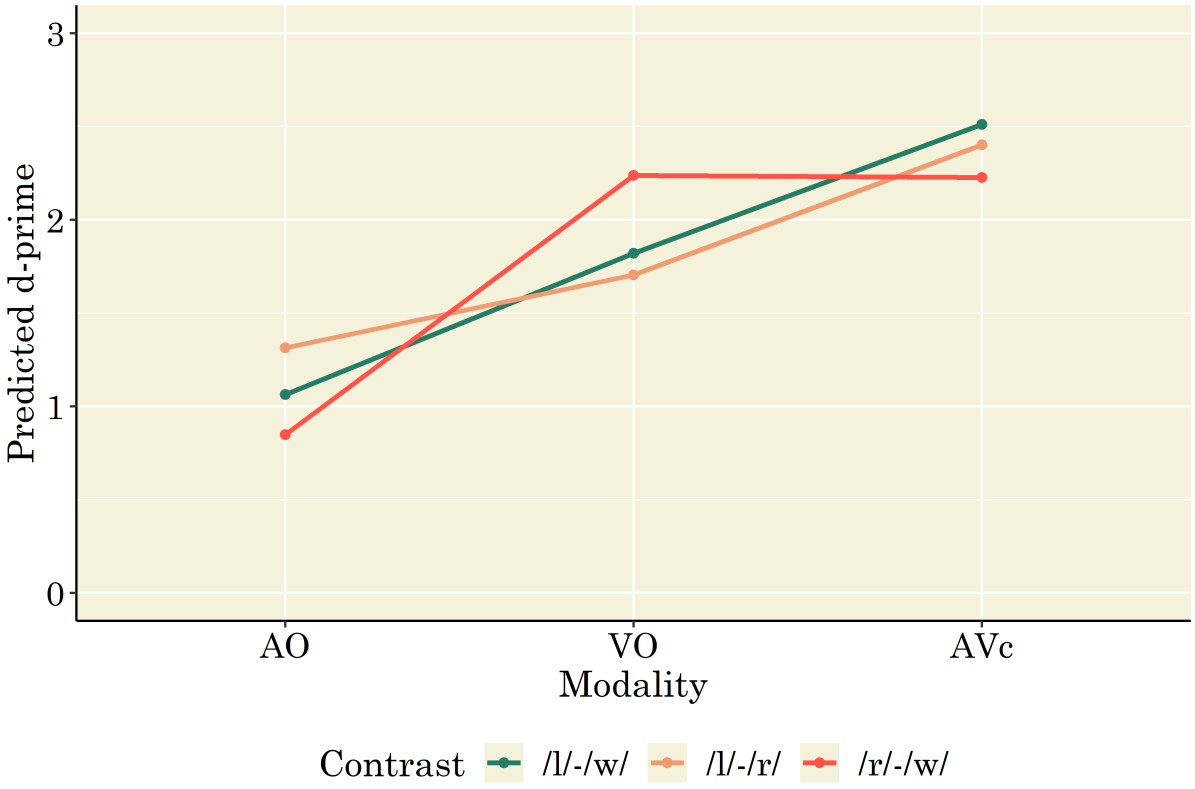



**Figure 6.11:** *Predicted sensitivity to* /l/-/r/, /l/-/w/ *and* /r/-/w/ *contrasts in each modality from a linear-mixed effects regression model.*

predicts the $d'$ values to be much higher than zero in the auditory-only modality in all three contrasts. The model predicts no significant difference in sensitivity to auditory cues between the three contrasts. Like the previous model, this one also suggests that /r/ has a labial gesture. If /r/ had no visible labial cue, we would expect the perception of /r/-/l/ to be significantly worse than that of /w/-/l/, which is not the case. Indeed, we observe no significant difference in sensitivity between /l/-/r/ and /l/-/w/ in the visual-only modality. These results indicate that subjects are sensitive to the difference in lip postures between /r/ and /l/, allowing us to conclude that /r/ indeed has a visible labial gesture.

As for the difference in lip postures between /r/ and /w/, according to our regression model, while the /r/-/w/ contrast in the auditory-only modality has the lowest predicted $d'$ of all the contexts under study, sensitivity to the /r/-/w/ contrast is significantly higher when only the visual cue is presented. The /r/-/w/ contrast, in fact, has the highest predicted $d'$ of all three contrasts in the visual-only modality. The regression model predicts sensitivity to all three contrasts to increase from the auditory-only to the visual-only one, but this increase fails to reach significance in the /l/-/r/ contrast. This is perhaps due to the high predicted sensitivity to the /l/-/r/ contrast in the auditory-only modality relative to the other two contrasts, resulting in a comparatively smaller increase from AO to VO. Sensitivity to all three contrasts increased from the auditory-only to the visual-only modality, but the cumulative benefit of the audio-visual modality is only observed for the contrasts with /l/, i.e., /l/-/w/ and /l/-/r/. Contrary to the other two contrasts under study, for the /r/-/w/ contrast, no significant benefit was obtained from presenting an auditory stimulus alongside the visual one. In other words, there is no significant difference in sensitivity to the /r/-/w/ contrast between the visual-only and the audio-visual modality. These results therefore suggest that the visual modality provides the highest sensitivity to the /r/-/w/ contrast. It would seem then that the visual cue of the lips may be more perceptually salient than the acoustic one (at least when presented in noise) for the /r/-/w/ contrast in Anglo-English.

### 6.3.4 *Perception of incongruent audio-visual trials*

Thus far, the results from the unimodal and congruous audio-visual stimuli indicate that /r/ has a visible labial gesture, which is perceptibly distinct from that of /w/. We also observed that the visual cues for /r/ and /w/ may actually be even more salient than the acoustic ones, at least when the auditory signal has been masked in noise. We will now turn our attention to the incongruent audio-visual stimuli. As Werker et al. (1992) indicate, in bimodal speech perception, when the visible articulation unambiguously specifies a particular place of articulation, visual capture can be anticipated. A high rate of visual capture would therefore provide further evidence to suggest that the labial postures for /r/ and /w/ are perceptually unambiguous and visually salient in Anglo-English.

We present results from incongruous audio-visual /r/-/w/ and /s/-/th/ trials. /s/-/th/ trials were included as controls because they allow for relatively straightforward predictions. The dental articulation of [θ] and [ð] should be relatively visible and therefore unambiguous, contrary to [s] whose primary articulation occurs inside the mouth and should not be visible to listeners. As a result, we predict visual capture to occur in incongruous audio-visual /s/-/th/ pairs. In contrast, incongruous /th/-/s/ audio-visual pairings should not induce visual capture. We hypothesise that /w/ and /r/ have a visible labial component which differs between the two phonemes, and as a result, visual capture should be possible in both /r/-/w/ and /w/-/r/ incongruous audio-visual pairings. We therefore predict that /r/, /w/ and /th/ visual stimuli will induce more visual responses than /s/.

Participants responded to 48 trials in which auditory /s/ was dubbed over visual /th/, auditory /th/ was dubbed over visual /s/, auditory /r/ was dubbed over visual /w/ and auditory /w/ was dubbed over visual /r/. Table 6.12 presents stimulus-response confusion matrices for all four incongruous audio-visual pairings. As with the unimodal and congruous trials, some incongruous trials were left unanswered. 29 out of the 39 subjects did not respond to all incongruous trials, averaging at 2.68 ± 2.39 unanswered trials per subject. The highest number of unanswered trials in any participant was 12/48. As with the analysis of unimodal and congruous trials, the 74 unanswered trials were excluded from subsequent analyses.

| | Visual stimulus | | | |
|---|---|---|---|---|
| | /s/ | /th/ | /w/ | /r/ |
| Visual response | 105 | 265 | 283 | 375 |
| Auditory response | 349 | 182 | 165 | 73 |
| No response | 14 | 20 | 20 | 20 |

**Table 6.12:** *Confusion matrices presenting responses to incongruent audio-visual trials.*

Figure 6.12 presents plots of the proportion of auditory and visual responses induced by the four incongruous contexts. As predicted, the confusion matrices in Table 6.12 and the proportions of responses presented in Figure 6.12 indicate that higher rates of visual responses arose from /th/, /w/ and /r/ visual stimuli than /s/. We observe an extremely large proportion of visual capture in the case of auditory /w/ dubbed with visual /r/ at 83.7%. The opposite context, i.e., auditory /r/ paired with visual /w/, resulted in a smaller proportion of visual responses (63.1%), although visual responses were still more frequent than auditory ones in this context.

To assess whether the proportion of visual capture significantly differs across the four incongruous audio-visual contexts, a generalised linear mixed-effects was implemented predicting response as the binary outcome variable (visual versus auditory). Fixed effects included visual stimulus (/s/, /th/, /w/, /r/), subject sex, age, origin and hearing score. Random intercepts were included for subjects and items. Likelihood ratio tests revealed that visual stimulus was a highly significant predictor ($\chi^2(3) = 43.20$, $p < .0001$). Subject sex was also significant ($\chi^2(1) = 4.51$, $p = 0.03$). However, none of the other fixed effects reached significance (Age: $\chi^2(1) = 0.20$, $p = 0.65$, Origin: $\chi^2(1) = 0.70$, $p = 0.40$, Hearing: $\chi^2(1) = 0.07$, $p = 0.79$). The output of the best-fitting model is presented in Table 6.13 and Figure 6.13 presents plots of the predicted probability of selecting a visual response for each incongruous audio-visual pair according to the model. Following our prediction, visual /s/ paired with auditory /th/ resulted in significantly fewer visual responses than the three other contexts. While no significant difference is observed between /w/ and /th/ visual stimuli, the model predicts that /r/ induces significantly more visual responses than /w/. /r/ also has the highest predicted probability of

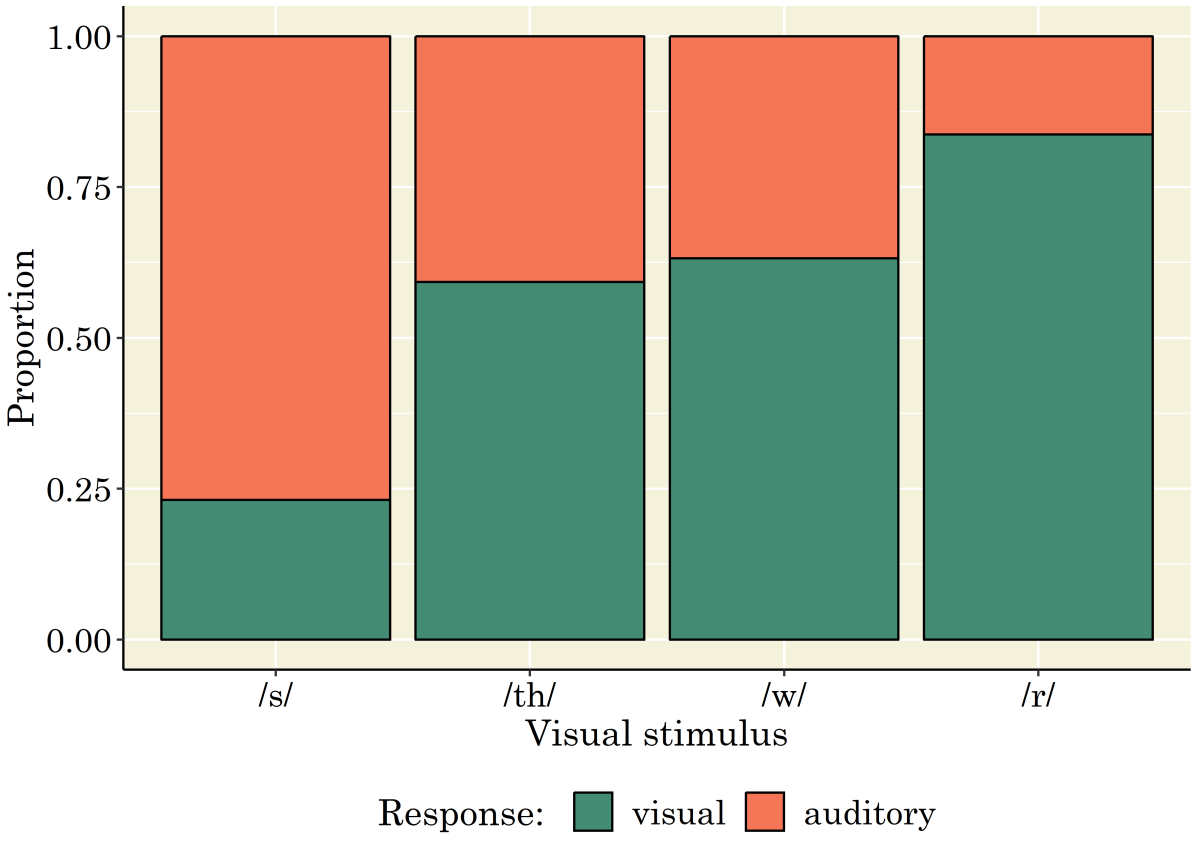



**Figure 6.12:** *Proportion of auditory and visual responses in incongruous audio-visual trials.*

visual capture of the four visual stimulus contexts (/s/ = 11.6%, /th/ = 56.1%, /w/ = 62.4%, /r/ = 88.1%).

| Predictor | Estimate (log-odds) | Std. Error | *t* value | *p* value |
|---|---|---|---|---|
| (Intercept) | 1.17 | 0.42 | 2.78 | $0.005^{**}$ |
| Visual stimulus /r/ | 1.50 | 0.49 | 3.05 | $0.003^{**}$ |
| Visual stimulus /s/ | -2.54 | 0.49 | -5.19 | $< .001^{***}$ |
| Visual stimulus /th/ | -0.26 | 0.48 | -0.55 | 0.59 |
| Sex male | -0.81 | 0.37 | -2.19 | $0.03^{*}$ |

*Visual response ~ Visual stimulus + Sex + (1|Subject) + (1|Item)*

**Table 6.13:** *Output of a generalised linear mixed-effects model predicting the probability of a visual response in incongruous audio-visual stimuli. The intercept represents trials containing visual /w/ perceived by female subjects.*

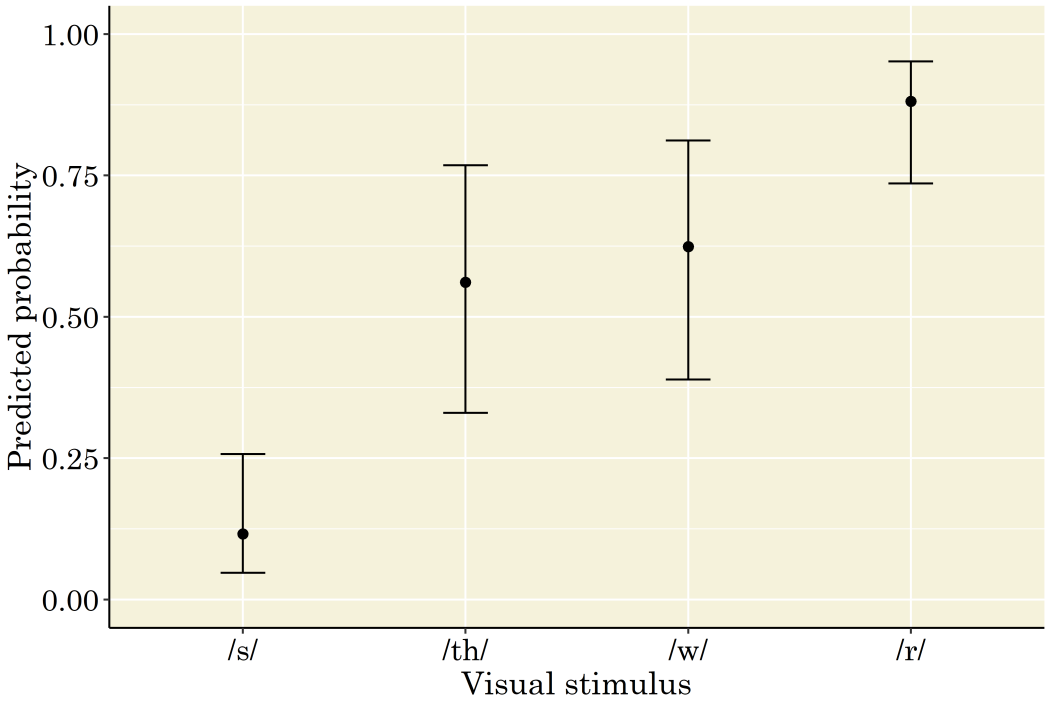



**Figure 6.13:** *Predicted probability of selecting a visual response in incongruous audio-visual trials from a generalised linear mixed-effects model.*

### 6.3.5 *Summary of results*

Putting together the various analyses presented in this section, the following results emerge. Firstly, we can confirm that the addition of pink noise did not completely mask the acoustic cues which distinguish /l/, /r/ and /w/ because sensitivity measures indicate that perceptual performance was not at all random in the auditory-only modality for all three contrasts (/l/-/r/, /l/-/w/, /r/-/w/). Indeed, the proportion of correctly identified stimuli in the auditory-only modality was around 73% on average, far above chance level. The lowest proportion of correctly identified stimuli in this auditory-only modality involved the /r/-/w/ contrast at 69% average accuracy. When presented with /w/ auditory-only stimuli, subjects actually reported perceiving /r/ more than /w/. This particular context resulted in a bias for /r/ responses. However, this response bias for /r/ in the /r/-/w/ contrast did not extend to the other two modalities in which visual speech cues were presented. Furthermore, while the auditory-only modality was the most challenging in this particular context, perceptual sensitivity to the /r/-/w/ contrast significantly increased with the presence of visual cues of the lips. Indeed, the average proportion of correct responses in the visual-only modality for /r/-/w/ is extremely high at 96%. These results therefore indicate that Anglo-English subjects are capable of distinguishing between the lip postures of /r/ and /w/ and that the visual cues of the lips may actually be more perceptually salient than the acoustic ones – at least when the auditory signal has been masked with noise. The results from unimodal and congruous audio-visual stimuli allow us to confirm that the lip posture for /r/ is visibly different from that of /l/. As it is generally agreed that onset /l/ is not labialised,[11] these results confirm that /r/ is produced with labialisation. We can conclude that as listeners are generally able to distinguish between /r/ and /w/ simply by visualising the lips, labialisation in /r/ and in /w/ is not implemented in the same manner. Results from the incongruous audio-visual trials provide further evidence to suggest that labialisation for /r/ is perceptually unambiguous. When subjects are asked to identify tokens in which auditory /w/ is dubbed with visual /r/, we observe a strong influence of the visual input with over 83% of all responses being the visual /r/ rather than the auditory /w/ cue. Interestingly, the rate of

[11] Inspection of lip camera data confirmed the absence of a visible lip posture for /l/.

visual responses is significantly smaller for mismatched auditory /r/ with visual /w/ tokens, indicating that labialisation in /w/ is more ambiguous than that of /r/, which will be discussed in greater detail in Section 6.4.

## 6.4 Discussion

In Experiment 2, it was proposed that a specific labial posture has evolved for Anglo-English /r/ due to increased exposure to labiodental variants. These non-lingual variants do not generate the low third formant frequency typically associated with post-alveolar articulations of /r/ and as a result, share more similar acoustic properties with /w/ than with lingual /r/, which may cause perceptual ambiguity between the two phonemes. The evolution of a specific labial gesture to accompany lingual productions of /r/ was thus related to the necessity to enhance the perceptual saliency of /r/ *acoustically*. In speakers who continue to produce /r/ with a specified tongue gesture, we argued that the acoustic effect of the typical labial posture we observed in these speakers may prevent over-lowering of F2, which could now be the principal acoustic cue that distinguishes /r/ from /w/ in Anglo-English, as proposed by Dalcher et al. (2008). While we focused on the acoustic consequences of the labial gestures in /r/ and /w/ in Experiment 2, the present study provides evidence that labialisation enhances the perceptual saliency of /r/ versus /w/ *visually*.

First of all, our results confirmed that /r/, like /w/, has a visually detectable labial gesture. Perceptual sensitivity to the /w/-/l/ and the /r/-/l/ contrast did not significantly differ in the visual-only modality. If /r/ is not produced with a visible labial gesture, we would expect the identification of /r/ versus /l/, which is not labialised, to pose a challenge in the visual-only condition, which was certainly not the case. While this is an important finding in itself, the major point to emerge from the present perception study is that /r/ and /w/ have visibly distinct labial postures which Anglo-English observers use as phonetic cues in their perception of the contrast. Consequently, we propose that /r/ and /w/ require separate viseme mappings in Anglo-English.

Seeing the bottom half of the speaker's face not only enhances auditory perception of the /r/-/w/ contrast (i.e., it results in visual enhancement), but the visual cues provided by the speaker's lips may actually be more informative than the auditory ones, at least when the auditory cues are masked in noise. Although one could argue that the presence of pink noise in the auditory signal may have hindered the perception of /r/ and thus be the cause of the observed disparity between auditory-only and visual-only perception, our statistical model predicts the proportion of correctly identifying /r/ in the auditory-only modality to be well above chance. Furthermore, /l/ stimuli in the auditory-only condition were predicted to have a high degree of identification accuracy at 86.1%, indicating that participants could still pick up on auditory cues more generally, despite the adverse listening conditions. We point out, however, that pink noise does not necessarily affect all acoustic cues equally. For example, Adachi et al. (2006) compared the effect of varying the SNR in auditory-only perception of the /r/-/l/, /b/-/v/, and /s/-/th/ contrasts in American English listeners. They observed that /r/-/l/ was more tolerant to noise than the other two contrasts, which the authors equated to the impact of noise on the different acoustic properties associated with the phonetic realisations of these phonemes. Productions of /b/, /v/, /s/, and /th/ form aperiodic sounds with broadband spectra. /r/ and /l/, on the other hand, produce sonorant sounds, which by definition, create periodic noise. They also have a narrow spectral peak at a comparatively lower frequency (between 1 kHz-3 kHz). Adachi et al. (2006) therefore suggested that sonorants may be more tolerant to noise than fricatives and stops. However, given that /w/, like /r/ and /l/, is also sonorant, Adachi et al.'s account cannot explain the differences we observed in identification accuracy between /l/ and /w/ and /r/ and /w/ in the auditory-only modality.

Despite the potential limitations caused by the addition of noise, our results suggest that the phonetic cues provided by the visual modality may be comparatively more informative than the ones provided by the auditory one. In our review of the literature on audio-visual speech perception presented in Chapter 1, we suggested that the highest perceptual advantage from visual speech input still requires a certain degree of auditory input. Previous studies have also ascertained that intelligibility is substantially greater in audio-visual than in auditory-only

and visual-only speech perception combined. Interestingly, our results on the perception of the /r/-/w/ contrast do not follow these patterns. For one thing, sensitivity to the /r/-/w/ contrast was significantly greater in the visual-only than in the auditory-only modality. But crucially, we observed no significant difference in sensitivity to the /r/-/w/ contrast between the visual-only and the audio-visual modality, contrary to the contrasts involving /l/, i.e., /r/-/l/ and /w/-/l/, and to results from previous studies on audio-visual speech perception more generally. For the /r/-/w/ contrast, the perceptual advantage from visual speech does not require auditory input whatsoever, which leads us to conclude that the visual cues may actually be more informative (or at least less ambiguous) than the auditory ones. Results from the incongruous audio-visual modality further support this proposal. High rates of visual capture, particularly in the trials containing visual /r/ paired with auditory /w/, suggest that the labial posture for /r/ is not only unambiguous with respect to that of /w/, but that Anglo-English speakers may weigh visual cues more than auditory ones in their perceptual categorisation of /r/ and /w/, particularly when the auditory signal is masked in noise. Put together, the findings from the present study allow us to support Hypothesis 7: Perceptual sensitivity to the /r/-/w/ contrast is enhanced by visual cues of the lips in Anglo-English. In actual fact, our findings propel us to take this hypothesis one step further; we propose that seeing the speaker's lips provides a highly informative phonetic cue for the /r/-/w/ contrast, one that may even override auditory speech perception.

### *6.4.1 Implications for sound change*

The heightened sensitivity to visual phonetic cues in the perceptual categorisation of /r/ and /w/ in Anglo-English speaker/observers invites the question as to why this heightened sensitivity to visual cues might exist, given the fact that audition is consistently defined as the primary mode of communication in spoken languages. In the remainder of this section, we will propose answers to this question, drawing on existing theories of speech perception and perception-based accounts of sound change. The main premise of our argument is that Anglo-English /r/ now has a specific labial posture, which is visibly distinct from that of

/w/. We will argue that the evolution of a generalised labial posture for /r/ relates to the necessity to maximise the distinctiveness of the /r/-/w/ contrast, which is weakened by increased exposure to non-lingual labiodental /r/ variants. The development of the specific labial gesture for /r/ thus avoids misperception. We will define three scenarios which we propose catalysed the evolution of a generalised labial gesture for Anglo-English /r/, including perceptual compensation, hypercorrection and audio-visual enhancement.

**Scenario 1: Perceptual compensation for labiodental /r/**

As described in Chapter 1 (notably in Section 1.5.2, page 27), Ohala's perception-oriented account of sound change (e.g., Ohala, 1981) proposes that sound change stems from the misperception of the acoustic signal by the listener. According to this view, phonetic variation is largely predictable. When the phonetically experienced listener encounters variation in the acoustic speech signal, the listener may do one of three things:

1. Factor out predictable phonetic variation and successfully reconstruct the form intended by the speaker, preventing misperception (i.e., *perceptual compensation*).
2. Take the acoustic signal at face value and fail to correct for phonetic variation, resulting in misperception (i.e., *hypocorrection).*
3. Make an erroneous correction of the acoustic signal, resulting in misperception (i.e., *hypercorrection*).

One of the key factors at play in Ohala's model is therefore the listener's phonetic experience. Without phonetic experience, the listener would simply be forced to take the acoustic signal at face value (i.e., via hypocorrection). We propose that increased phonetic experience may allow Anglo-English speakers to correctly reconstruct labiodental productions of /r/ as /r/ and not as /w/. As we have indicated throughout this thesis, Anglo-English listeners are regularly confronted with phonetic variation for /r/. While tongue shapes for post-alveolar productions of /r/ vary from bunched to retroflex, the acoustic output of these pronunciation variants remains comparatively stable, as detailed in Chapter 4. However, articulations without

a specified lingual component, such as labiodental ones (e.g., [ʋ]), do not produce the same acoustic output as lingual articulations. As suggested in Dalcher et al. (2008), labiodental /ɹ/ with its high third formant frequency, may share more acoustic properties with [w] than with lingual /ɹ/. Indeed, perceptual confusion between [ʋ] and [w] is widely described in the literature. For example, Foulkes and Docherty (2000) reviewed historical evidence of labiodental /ɹ/ and observed a tendency for it to be represented orthographically as 'w' both in traditional literature, including works by George Orwell and Charles Dickens, and in more contemporary media, such as in television and advertising. In addition, it is often reported that children acquiring English substitute word-initial /ɹ/ with [w], although experimental evidence suggests that children do produce different phonetic realisations of /ɹ/ and /w/ (Dalston, 1975; Kuehn & Tomblin, 1977). It may be that children actually substitute lingual /ɹ/ for labiodental /ɹ/, and its acoustic proximity to [w] results in it being erroneously classified as [w] by adults. Indeed, Knight et al. (2007) took acoustic measures from a child acquiring SSBE, which suggested that the transition from [w]-like articulations to more adult-like articulations of /ɹ/ includes an intermediary labiodental variant. The steady mastery of acoustics with a gradual raising of F2 and a lowering of F3 detailed by Knight et al. (2007) has also been observed in children acquiring American English (S. Lee et al., 1999; McGowan et al., 2004).

As labiodental variants are rapidly gaining currency across England, phonetic experience of such variants must also be on the rise. Incidentally, increased exposure may explain why labiodental /ɹ/ is becoming less stigmatised (see Foulkes & Docherty, 2000, for a detailed review of changing sociolinguistic perceptions of [ʋ]). Increased phonetic experience may allow listeners to factor out the acoustic variation resulting from non-standard labiodental variants and reconstruct the form intended by the speaker as /ɹ/. This scenario may be schematised based on Ohala's depiction of *perceptual compensation*, which we present in Figure 6.14. Perceptual compensation is often associated with the listener's ability to factor out acoustic variation caused by coarticulation, i.e., in studies on compensation for coarticulation (e.g., Beddor, Harnsberger, & Lindemann, 2002; Beddor & Krakow, 1999; Harrington et al., 2008; Mann & Repp, 1980, among many others). However, we see no reason why perceptual

compensation might not be extended to /r/ pronunciation variants. As compensation requires phonetic experience in the listener, we predict that listeners who lack exposure to labiodental /r/, for example American English speakers, would be less likely to reconstruct [ʋ] as /r/. Given its acoustic proximity to [w], we predict that American English listeners would likely interpret [ʋ] as /w/. This scenario involving inexperienced listeners would therefore correspond to one of *hypocorrection*, as proposed by Ohala.

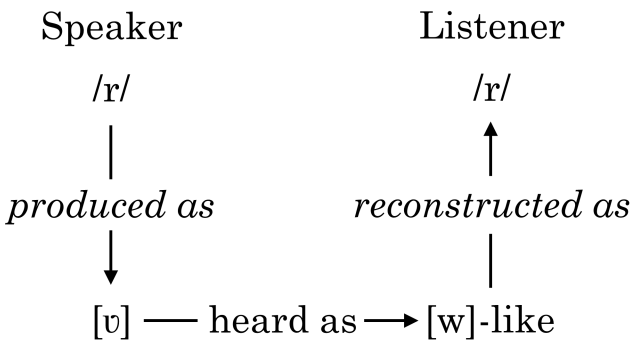



**Figure 6.14:** *Schematisation of perceptual compensation for labiodental* /r/ *in Anglo-English listeners, based on Ohala's perception-oriented account of sound change.*

While the present perception study did not examine the perception of labiodental /r/, the suggestion that phonetic experience of /r/ variation may influence the perception of the Anglo-English /r/-/w/ contrast is further strengthened by the significant effect of speaker origin in our dataset. In unimodal and congruous audio-visual perception, participants who spent their childhood in England were more sensitive to contrasts and attained significantly higher accuracy than those who grew up abroad. However, we stress that the dataset was not balanced for participant origin and only contained data from three participants who did not grow up in England. This result therefore requires further investigation. Similarly, we may expect participant age to influence perception in a similar manner, but age as a predictor did not reach significance in any of our statistical models, although a non-significant result does of course not necessarily mean that there is no effect. We note that this experiment was not

designed to test predictions about age and so, like participant origin as a factor, our data were not stratified for age. The effect of age therefore could be the goal of another future study.

Parenthetically, intersubject variability in our perception results also came through in the significant effect of participant sex. Women were more sensitive to contrasts and performed more accurately than men in unimodal and in congruous audio-visual perception. Women were also more likely to select a visual response than men in incongruous audio-visual trials. As we described in Chapter 1, previous studies have made the same observation and the reported female advantage in utilising visual cues from speech has been related to differences in brain activity between men and women (Desjardins & Werker, 2004).

**Scenario 2: Hypercorrection of [w] results in auditory /r/-bias**

An unexpected result emerged in auditory-only perception of the /r/-/w/ contrast in the present study. When Anglo-English speaking participants were presented with auditory productions of word-initial /w/ and were asked to choose between words beginning with 'w' and 'r', more 'r' responses than 'w' responses were selected. This resulted in a bias for /r/ in the auditory perception of the /r/-/w/ contrast. We propose that high exposure to labiodental variants in Anglo-English means that English listeners have to tolerate a high degree of acoustic variation for /r/, which may actually be detrimental to the perception of /w/. As the previous scenario indicated, Anglo-English listeners may have to regularly factor out acoustic variation in their perception of /r/, but their phonetic experience of /r/ variation may result in erroneous corrections of /w/ productions. Given the acoustic similarity between [ʊ] and [w], a speaker's /w/ productions may be incorrectly classified as /r/ by the listener, which would account for the observed /r/ bias in the identification of /w/-/r/ target-distractor pairs in the auditory-only modality. This is an example of what Ohala defines as *hypercorrection*, and is schematised in Figure 6.15. Again, phonetic experience of labiodental /r/ is a crucial element for this schematisation to apply. We would not expect hypercorrection of canonical productions of /w/ to be reconstructed as /r/ in English listeners who have not been heavily exposed to labiodental /r/.

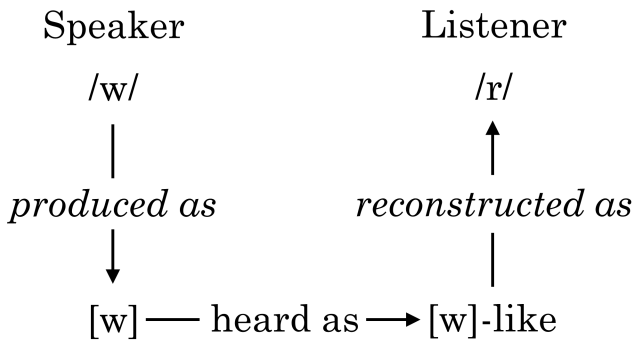



**Figure 6.15:** *[Schematisation of hypercorrection of* /w/ *to* /r/ *in Anglo-English listeners, based on Ohala's perception-oriented account of sound change.*

However, we point out that an alternative account for the bias observed for /r/ could involve word frequency. A higher frequency of onset /r/ than onset /w/ in English would also explain why listeners tend to select 'r' rather than 'w' responses when presented with auditory /w/ in noise. Consequently, frequency counts are still necessary in order to verify that the observed /r/ bias is more likely the result of hypercorrection and not simply due to higher frequency word-onset (and perhaps syllable-onset) /r/ than /w/. Nevertheless, higher frequency onset /r/ would not account for the lack of an /r/ bias observed in the visual modalities (VO and AVc), which we will consider in our third and final perception scenario.

**Scenario 3: Visual cues prevent misperception-based sound change**

Not only did the addition of visual cues prevent /r/ bias in the auditory perception of the /r/-/w/ contrast, but perceptual sensitivity to the contrast significantly increased when participants could see the speaker. We have suggested that the visual phonetic cues for the /r/-/w/ contrast may actually be more informative than the auditory ones, given the fact that the perceptual advantage from visual cues did not require auditory input in any form. Similarly, incongruous audio-visual stimuli generally resulted in Anglo-English speaker/observers being visually captured. Given these results, the two scenarios we previously proposed involving

hypercorrection and perceptual compensation can no longer apply when listeners have access to visual speech cues.

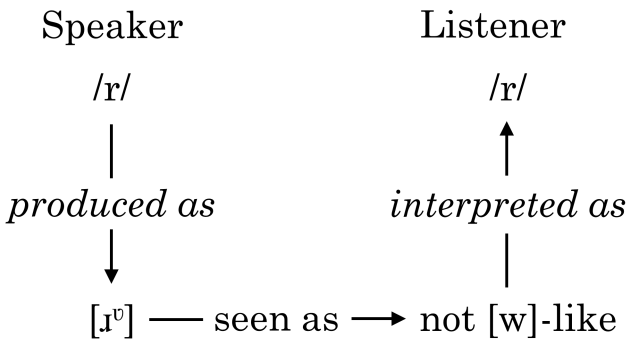
In Chapter 5, we showed that labialisation is implemented differently for /r/ and /w/; /r/ is produced with a more labiodental-like posture than /w/ (i.e., /r/ presents vertical labialisation, while /w/ has horizontal labialisation). The results from the present study indicate that Anglo-English speaker/observers are sensitive to the different labial configurations for /r/ and /w/, which may affect how the contrast is perceived. We propose that visual cues provide perceptually salient phonetic information concerning the identity of the phoneme in question, which may even override auditory perception of the /r/-/w/ contrast. When an Anglo-English listener is presented with a lip posture that is visibly not [w]-like, the listener will likely interpret that production as /r/ and not as /w/. This scenario is presented in Figure 6.16a using the same format as in the previous two scenarios, which were inspired by Ohala's perception-based account of sound change. Although Ohala's approach focuses entirely on auditory perception, it is now widely accepted that speech perception is multimodal. Consequently, we propose to extend Ohala's perceptual account of sound change to include visual cues, which better reflects the reality. We note that in Figure 6.16a, we use the diacritic [ᵘ] to denote the labiodental-like lip posture caused by vertical labialisation, which we associate with Anglo-English productions of /r/. Although we have used the phonetic symbol [ɹ] to refer to bunched tongue shapes previously in this thesis, in this instance, we stress that our use of [ɹ] encompasses all possible tongue shapes for the post-alveolar approximant. We have indicated that the auditory cues for /r/ and /w/ may be perceptually ambiguous in Anglo-English speaking listeners, given the presence of labiodental /r/. Our results demonstrate that visual cues allow speaker/observers to better disambiguate the contrast, which is reflected in the schematisation in Figure 6.16a.

Figure 6.16b presents a schematisation of the visual perception of /w/. When the listener sees a [w]-like visual cue (presented in Figure 6.16b with the diacritic [ʷ]), the listener will likely interpret that realisation as /w/ and not as /r/, given their phonetic knowledge of the visually distinct labial cues for /r/ and /w/. If we compare this scenario with the hypercorrection scenario schematised in Figure 6.15, we observe that the presence of visual cues prevents the

hypercorrection of [w] to /r/ from occurring. Importantly, by preventing hypercorrection, visual cues may avert potential misperception-based sound change. Ohala proposes that sound change may arise due to misperception in the listener when the listener turns speaker. For example, we may imagine an extension of Scenario 2 in which hypercorrection results in more [w]-like realisations of /r/ being produced when the listener turns speaker because the listener increasingly associates [w]-like productions with /r/. However, the presence of visual cues may prevent such a scenario from taking place because the visual cues from the lips seem to provide less ambiguous phonetic cues than the auditory ones, thus allowing the listener to correctly interpret visible productions of /r/ and /w/ as /r/ and /w/.

While we would argue that the /r/-typical posture is unequivocally associated with /r/, the same cannot necessarily be said for the posture we associate with /w/. It was proposed in Chapter 5 that the acoustic effect of the vertical labialisation observed in /r/ productions was to enhance F3 lowering by extending the front cavity with lip protrusion, all the while maintaining a maximally high F2. However, an extension of the front cavity for /r/ via increased lip protrusion could also be accomplished with the horizontal labialisation we typically observed in /w/ productions, although this would likely result in a lower frequency F2 for /r/ than in realisations with vertical labialisation. As both lip postures may be produced with lip protrusion, both of them could technically enhance F3 lowering for /r/. In contrast, as we indicated in Chapter 5, articulatory-acoustic models generally converge on the suggestion that in the case of a backed tongue constriction, such as the one produced for [w], in order to achieve a minimally low F2, a labial constriction produced with close lip rounding (i.e., horizontal labialisation) is vital (e.g. Fant, 1960; Stevens, 1998). It is therefore highly improbable that unrounded productions of [w] would naturally occur without having a detrimental effect on F2. Vertical labialisation thus seems entirely incompatible with productions of [w] with a canonically low F2.

Consequently, while a visual cue presenting labiodental-like vertical labialisation can only accompany productions of /r/, horizontal labialisation may perhaps naturally occur in productions of both /w/ and /r/. The results from the incongruous audio-visual trials further




**(a)** *Visual perception of* /r/

*Figure omitted due to missing copyright permission*

*(Elément sous droit, diffusion non autorisée)*

**(b)** *Visual perception of* /w/

**Figure 6.16:** *Schematisation of visual perception of* **(a)** /r/ *and* **(b)** /w/ *in Anglo-English, using a similar format to the sound change scenarios proposed by Ohala.*

support the suggestion that the labial posture for /r/ is unequivocally associated with /r/, while the labial posture for /w/ is more ambiguous. When a visual cue of /r/ is paired with an auditory cue of /w/, the average proportion of selecting a visual response was predicted by a generalised linear mixed-effects model to be extremely high at nearly 90%, which suggests that the visual cue of the lips for /r/ is unambiguous with respect to /w/. In opposing incongruous audio-visual pairings, where visual /w/ is paired with auditory /r/, significantly fewer visual responses occur, with an average proportion of visual responses predicted at just over 60%. Visual capture is therefore more likely to take place when native Anglo-English-speaking participants are presented with the visual cue of /r/ rather than the visual cue of /w/, which indicates that the labial posture for /w/ is more perceptually ambiguous than that of /r/. We may predict, however, that over time, the labial cues for /r/ and /w/ will become more unambiguous, as exposure to productions of /r/ presenting the /r/-typical labiodental-like posture increases.

We have argued throughout this thesis that the labial gesture we have observed for /r/ is typical of Anglo-English and has evolved due to the necessity to increase the perceptual salience of /r/, as a result of high exposure to non-lingual /r/ in this variety. This assertion infers that /r/ was not always produced with a unified labial gesture in the past, nor is it produced with one posture in other varieties of English. Indeed, in their recent articulatory study on American English /r/, B. J. Smith et al. (2019) described a labial posture for /r/ that is much more variable than the one we have observed in Anglo-English, as they found both horizontal labialisation and vertical labialisation, according to our definitions. In the literature on viseme mappings for /r/ and /w/ presented in Section 6.1.2, we noted that Fisher (1968) found that in American English speaker/observers, /r/ as a stimulus was significantly confused with /w/, but /w/ was not confused with /r/. The present study found the opposite trend. In incongruous audio-visual trials, seeing /r/ resulted in more visual capture than seeing /w/, which indicates that while /w/ as a visual stimulus may be confused with /r/, visual /r/ is not confused with /w/. This is interesting because it indicates that there may be distinct differences in the visual perception of /r/ and /w/ in the two varieties of English, although the Fisher

(1968) study is now admittedly very dated. It may be that the lips are less perceptually salient in the American English contrast. As Dalcher et al. (2008) suggested, the /ɹ/-/w/ contrast is still defined acoustically by the frequency of F3 in American English (and not F2), which we predict allows speakers more freedom in their implementation of labialisation. Anglo-English speakers may well have phonetic experience of /ɹ/ being produced with labial postures that are different to the typical labiodental-like one we have defined, both in other varieties of English and in past Anglo-English productions. As a result, combining an auditory cue of /ɹ/ with a non-typical labial cue, such as the one produced for /w/, would not necessarily result in a listener interpreting such a production as /w/, but may be classified as /ɹ/ with a non-typical lip posture. Furthermore, given the /ɹ/ bias which results from auditory perception of the contrast, we may imagine that speaker/observers may still perceive /ɹ/ in this context. This scenario may therefore account for the difference in visual capture rates we observed between /ɹ/ and /w/ visual cues in the perception of incongruous audio-visual /ɹ/-/w/ trials.

**Summary**

Scenarios 1 and 2 allowed us to account for the perceptual ambiguity observed in the auditory perception of /ɹ/ and /w/. Despite the acoustic similarity between [ʊ] and [w], increased exposure to the former enables Anglo-English-speaking listeners to reconstruct [ʊ] realisations as /ɹ/ and not as /w/ via perceptual compensation (presented in Scenario 1). However, increased phonetic knowledge of the acoustic resemblance of [ʊ] and [w] may also result in hypercorrection of /w/ productions to /ɹ/ in listeners, which results in a bias for /ɹ/ in the auditory perception of the /ɹ/-/w/ contrast (presented in Scenario 2). Hypercorrection is problematic because it may spark misperception-based sound change with [w]-like productions being increasingly associated with /ɹ/ rather than /w/, which could result in listeners producing more [w]-like realisations of /ɹ/ when it is their turn to speak. However, when the listener is able to see the speaker's lips, hypercorrection, and therefore misperception-based sound change, is less likely to occur for the /ɹ/-/w/ contrast (Scenario 3). We propose that this is because /ɹ/ now has a generalised labial posture in Anglo-English, which is visibly

different from that of /w/. The clear visible distinction between the labial gestures for /r/ and /w/ provide perceivers with perceptually salient phonetic cues, which enable them to better disambiguate the contrast than when they have access to auditory speech cues alone. In actual fact, the visual cues seem to provide the least ambiguous phonetic information of the two modalities.

## 6.5 CHAPTER CONCLUSION

By considering the audio-visual perception of the /r/-/w/ contrast in Anglo-English, we have shown that auditory perception of the contrast in noise is not only enhanced by seeing the speaker's lips, but that visual speech cues may provide more reliable phonetic information than the auditory ones. Exposure to non-lingual pronunciations of /r/ has forced Anglo-English listeners to tolerate such a high degree of acoustic variation that even a canonical production of /w/ may be reconstructed as /r/. In contrast, the visual cues for /r/ and /w/ are far less ambiguous. Participants are able to identify /r/ and /w/ from the visual cues alone (i.e., via lip reading) with an almost perfect degree of accuracy. Furthermore, by investigating incongruous audio-visual perception, we have shown that the visual cue of the lips for /r/ is salient enough to dominate a mismatched auditory cue of /w/ when presented in noise. We conclude that the results from the present study support our proposal from Experiment 2 that the labial posture accompanying lingual articulations of Anglo-English /r/ has evolved to be specific to /r/ in order to reinforce the phonological contrast with /w/. We predict that Englishes which generally lack non-lingual productions of /r/, such as American English, would not need to enhance the phonological contrast between /r/ and /w/, which would allow speakers more freedom when it comes to the implementation of labialisation for /r/. Finally, given the relative perceptual certainty of the visual cues from the lips contrary to the acoustic ones for Anglo-English /r/, one might predict a continued increase in the change from lingual to non-lingual labial articulations. Predicting future sound changes should be undertaken with caution, and so we conclude that for now, the articulation of Anglo-English /r/ remains to be seen – and not heard!

# General discussion and conclusions

# 7

The main goal of this thesis was to contribute to our understanding of the role of the secondary labial gesture in the production and perception of post-alveolar /r/ in Anglo-English. It was suggested that the lips may be particularly important for /r/ in this variety because non-lingual labiodental variants are rapidly gaining currency across England. While all of the speakers presented in this thesis had an observable lingual gesture for /r/, which varied in shape from curled up retroflex to tip down bunched in a similar way to the articulatory variation documented in other Englishes, our results indicate that the lips are used by speakers to enhance the discriminability of /r/ in both the auditory and the visual domains. We propose that the Anglo-English post-alveolar /r/ is produced with a specific labial posture, which allows speakers to increase the size of the anterior buccal cavity and therefore enhance the lowering of the third formant without having a significant impact on the second formant. Over-lowering of the second formant would result in perceptual uncertainty with /w/, particularly because increased exposure to high-F3 labiodental variants of /r/ may have resulted in a cue shift from F3 to F2 in the perception of the /r/-/w/ contrast in England, as proposed by Dalcher et al. (2008). This /r/ specific labial posture has consequences not only for the auditory perception of the /r/-/w/ contrast, but also for its visual perception. When Anglo-English listeners are

asked to identify canonical productions of /w/ (i.e., with a high F3 and a low F2) presented in noise, they actually respond with /r/ more often than /w/. This suggests that listeners have to tolerate such a high degree of variability for /r/ that even canonical productions of /w/ may be reconstructed as /r/. However, when presented with the accompanying visual cues, the bias for /r/ responses disappears and sensitivity to the contrast is significantly enhanced. In actual fact, perception results suggest that the visual cues of the lips may be more phonetically informative than the auditory ones. We thus conclude that Anglo-English speakers use increased labiality to enhance the auditory and visual perceptibility of the post-alveolar approximant /r/. Below, we summarise the results that led us to make these conclusions. We will then discuss the implications of these findings to the wider field, notably concerning the nature of speech perception and the role of visual cues in the evolution of sound systems. Finally, we will consider the limitations of this study as well as the directions future research may take.

## 7.1 MAIN FINDINGS

### *7.1.1 Tongue shapes for Anglo-English post-alveolar /r/ are variable*

In Experiment 1, we presented articulatory data from 24 Anglo-English speakers' productions of /r/. A variety of tongue shapes were observed ranging from curled up retroflex to tip down bunched. It was hypothesised that retroflex tongue shapes would be more common than bunched ones in Anglo-English given the results from Heyne et al. (2018) on non-rhotic New Zealand English. Heyne et al. (2018) speculated that as New Zealand English speakers rarely produce post-vocalic /r/, where bunching is heavily favoured, speakers are less likely to acquire bunched /r/ as an alternative articulation strategy if they have already mastered retroflexion. Our results from Anglo-English support this suggestion because we observed double the number of retroflex-only users than bunched-only. While some speakers use one tongue configuration exclusively in all vowel contexts, others present consistent but individual variation. This finding has also been observed in previous studies of /r/ in other varieties of English, such as American English. Mielke et al. (2016) observed that retroflexion is more

compatible in the context of open-back vowels as opposed to close-front ones and our results on Anglo-English follow the same pattern. Coarticulation with the following vowel occurs not only in relation to tongue shape, but tongue position is also affected. We observed that the lingual constriction is generally fronted when /r/ is followed by a front vowel, which appears to have acoustic consequences. While no significant differences in the frequency of the third formant were found for the different tongue shapes, significant differences were observed with regards to the following vowel. The lowest F3 values for /r/ were predicted by our statistical model to coincide with the backest vowels under study. These results therefore support the following hypotheses:

**Hypothesis 1** Anglo-English /r/ is produced with higher rates of retroflexion than in American English.

**Hypothesis 2** Tongue shapes for Anglo-English /r/ are affected by coarticulation with the following vowel.

**Hypothesis 3** Different tongue shapes for Anglo-English /r/ result in similar formant values – at least up to F3.

### *7.1.2 The lips enhance the auditory cues for Anglo-English /r/*

Our review of the existing literature on the articulation of English /r/ presented in Chapter 2 led us to hypothesise that lip protrusion may contribute to the lowering of the third formant (Hypothesis 4). It is generally agreed that F3 is the most salient acoustic cue of /r/. The low frequency F3 for /r/ has been associated with a large cavity volume in front of the palatal constriction, which includes the sublingual space. Multi-tube models as well as physical models of the vocal tract indicate that the lowest possible third formant values are generated by the largest possible front cavity volumes (e.g., Alwan et al., 1997; Espy-Wilson et al., 2000; Lindblom et al., 2010; Stevens, 1998). Increasing the size of the front cavity should therefore decrease F3. Possible strategies to increase the size of the front cavity include backing of the palatal constriction, increasing the size of the sublingual cavity through increased retroflexion,

or increasing the size of a lip protrusion channel in front of the palatal constriction. To test whether speakers use the latter strategy of increased lip protrusion, we elicited hyperarticulated productions of /r/. It was predicted that hyperarticulated productions of /r/ would result in even lower F3 values than those observed in non-hyperarticulated ones and that speakers may achieve these lower F3 values with increased lip protrusion. Our results indeed support these predictions. Hyperarticulated productions resulted in increased lip protrusion and lower F3 values than non-hyperarticulated ones. The results therefore support Hypothesis 4:

**Hypothesis 4** Lip protrusion contributes to the lowering of the third formant of /r/.

Our literature review also indicated that the various tongue shapes associated with the production of Anglo-English /r/ from curled up retroflex to tip down bunched (e.g., Delattre & Freeman, 1968) generate differing vocal tract dimensions, and yet the resulting acoustics are remarkably consistent. For example, retroflex tongue shapes have been found to produce larger front cavities than bunched shapes (Alwan et al., 1997), which we suggest may be due to the large sublingual space brought about by retroflexion. We predicted that bunched tongue shapes may compensate for their smaller front cavity with increased lip protrusion. Similar trading relations have been proposed for other speech segments, such as [u] (Perkell et al., 1993; Savariaux et al., 1995), and even for productions of English /r/ itself (Alwan et al., 1997; Guenther et al., 1999). Our results from both non-hyperarticulated and hyperarticulated productions of /r/ indeed suggest that bunched tongue shapes are produced with significantly more lip protrusion than retroflex ones. We suggested that retroflex users may increase the size of the front cavity via increased retroflexion, a strategy which would not be available to bunchers, and as a result, we predicted that bunched /r/ may result in higher rates of lip protrusion in hyperarticulation, which was indeed the case. These results therefore support Hypothesis 5:

**Hypothesis 5** A trading relation exists between the size of the sublingual space and the degree of lip protrusion which manifests itself through a negative correlation between the two.

We therefore concluded that by extending the front cavity, increased lip protrusion con-

tributes to the lowering of the third formant, the most salient acoustic cue for the post-alveolar approximant /r/, and therefore to its acoustic discriminability. Our results indicated that while increased lip protrusion lowered F3, F2 was not significantly affected. Given what we know about the effect of lip *rounding* on formant frequencies, particularly on F1 and F2 (as presented in Chapter 3), this result was somewhat unexpected. If increased lip rounding (i.e., involving a decrease in lip area) is a concomitant of increased lip protrusion, we would expect significant decreases in F2, which was not the case. We proposed that increased labiality may thus allow speakers to lower F3 for /r/ while maintaining a small distance between F2 and F3. Indeed, researchers have remarked on the close proximity of F3 to F2 for English /r/ (Dalston, 1975; Guenther et al., 1999; Lisker, 1957; O'Connor et al., 1957; Stevens, 1998). As we have discussed, labiodental variants of /r/ are becoming increasingly common across England. Docherty and Foulkes (2001) define a change in progress whereby the labial component of Anglo-English /r/ is 'retained at the cost of the lingual articulation' (p. 183), which we remarked implies that the labial gesture for /r/ is labiodental even when accompanied by a specified lingual posture. By comparing the lip posture of /r/ with that of /w/, which is unequivocally described as rounded, we aimed to verify this claim. We predicted that if lingual /r/ was produced with a labiodental component, the lip postures for /r/ and /w/ should differ considerably.

In Experiment 2, the lip postures for /r/ and /w/ were compared using the front lip camera data recorded in Experiment 1. A variety of methods were used to measure and analyse lip postures, including techniques from deep learning. Both hand measures and those obtained from a Convolutional Neural Network which automatically segmented the lip area from the rest of the image, indicated that the lip posture for /r/ and /w/ differ. The most robust differences occurred in the width and vertical position of the lips: the lips are wider and higher for /r/ than they are for /w/. We concluded that labialisation is implemented via horizontal labialisation for /w/ and vertical labialisation for /r/. We accounted this difference in lip postures to the increased pressure to maintain an auditory contrast between /r/ and /w/ beyond F3 caused by increased exposure to non-lingual (high F3) variants. Horizontal labialisation results in tightly closed lip rounding which is associated with F2 lowering. By avoiding this tight lip

rounding with vertical labialisation, Anglo-English speakers who still produce an observable lingual constriction for /r/ are able to generate low F3 values (including lip protrusion) without over-lowering F2. In Englishes where non-lingual variants of /r/ are less common, such as in American English, we predicted that speakers should be freer to vary their labial posture for /r/, which previous articulatory studies appear to suggest (e.g. B. J. Smith et al., 2019). Finally, we observed that this /r/ specific labial posture shares similar articulatory characteristics with labiodental articulations. In order to protrude the lips without contracting them at the corners, the bottom lip is extended up and outwards upwards towards the top teeth. We concluded that if the /r/ specific labial posture is visually salient, as proposed by Docherty and Foulkes (2001), we may predict a continued increase in non-lingual labiodental variants of /r/ in England, as the lingual component is dropped for the labial one. These results therefore support Hypothesis 6:

**Hypothesis 6** /r/ has a specific lip posture which differs from that of /w/ in Anglo-English.

### *7.1.3 The lips enhance the visual cues for Anglo-English* /r/

Docherty and Foulkes (2001) postulated that labiodentalisation may be a 'function of the heavy visual prominence of the labial gesture' (p. 183). To assess to what extent the labial component described in Experiment 2 is visually salient to Anglo-English native speakers, we conducted a perception study in the final experiment presented in this thesis, Experiment 3. We suggested that the evolution of a specific labial posture for /r/ may have occurred in order to enhance the perception of /r/ visually as well as auditorily, by increasing the phonetic contrast between /r/ and /w/. In Experiment 3, Anglo-English subjects were presented with auditory-only, visual-only, congruous audio-visual and incongruous audio-visual productions of /r/ and /w/ in noise produced by a native speaker of Anglo-English. Productions of /l/ were also included to act as a control. It was observed that subjects could distinguish /w/ from /l/ and /r/ from /l/ using the visual speech cues alone, suggesting that both /w/ and /r/ have a visible labial cue. Furthermore, perceptual sensitivity to the /r/-/w/ contrast significantly increased with the presence of visual cues of the lips. Correct identification of /r/ and /w/ tokens was significantly

higher in the audio-visual than in the auditory-only modality. This result allowed us to support the final hypothesis of this thesis:

**Hypothesis 7** Perceptual sensitivity to the /r/-/w/ contrast is enhanced by visual cues of the lips in Anglo-English.

In addition, the results from Experiment 3 allowed us to suggest that the visual cues may actually be more salient than the auditory ones for the /r/-/w/ contrast. In the visual-only modality, participants achieved extremely high identification accuracy for the /r/-/w/ contrast and there was no significant difference between visual-only and audio-visual perception, unlike in the contrasts with /l/. As a result, contrary to the results from many previous studies on audio-visual speech perception, the perceptual benefit from visual speech cues in the Anglo-English /r/-/w/ contrast does not require auditory input at all. We proposed that the visual cues may provide more robust and less ambiguous phonetic information than the auditory ones. Results from the incongruous audio-visual trials further strengthened this argument because high rates of visual capture occurred, particularly in the context of visual /r/ paired with auditory /w/.

Finally, we proposed that high exposure to acoustic variation for /r/, notably in the form of non-lingual labiodental variants, may allow listeners to reconstruct labiodental productions of /r/ as /r/ and not /w/, despite the acoustic similarities between [ʋ] and [w]. However, the increased phonetic experience of [ʋ] may result in [w] productions being erroneously reconstructed as /r/ by listeners. This hypercorrection of /w/ could catalyse a sound change towards more [w]-like productions of /r/ when the listener turns speaker, as the listener increasingly perceives [w]-like productions as /r/. However, hypercorrection, and therefore this potential misperception-based sound change, is less likely to occur when the listener has access to visual speech cues. The visually distinct lip postures for /r/ and /w/ productions allow listeners to better disambiguate the contrast than in auditory perception alone. We therefore tentatively predicted that future productions of /r/ will become increasingly labialised, given the perceptual weight of its visible labial cue.

### *7.1.4 Answering the research questions*

The results from the three studies presented in this thesis allow us to provide possible answers to our research questions:

1. Is the tip-up tongue shape typical of post-alveolar approximant /r/ in Anglo-English?

Articulatory data from 24 Anglo-English speakers provided by Ultrasound Tongue Imaging indicated that although a range of tongue shapes are used from curled up retroflex to tip down bunched for /r/, tongue shapes produced with the tip or the front of the tongue raised towards the post-alveolar region of the palate with or without curling back are much more common than tip down ones.

(a) Is tongue shape subject to coarticulation with the following vowel as in other varieties of English?

Yes. Retroflexion is more compatible with low back vowels than close-front ones, potentially because the tongue tip is freer to move. The palatal constriction undergoes fronting in the context of front vowels regardless of tongue shape.

2. How does lip protrusion contribute to the production of Anglo-English /r/?

We propose that lip protrusion extends the size of the front cavity and may thus allow speakers to maintain a relatively consistent front cavity volume across different tongue shapes and constriction locations, which generate consistent acoustic outputs, particularly with regards to F3.

(a) Can lip protrusion enhance F3 lowering?

Yes. The results from Experiment 1 suggest that increased lip protrusion is used in hyper-articulated productions of /r/, which result in lower F3 values than in non-hyperarticulated productions.

(b) Is there a relationship between the degree of lip protrusion and lingual articulation?

Yes. We found higher degrees of lip protrusion in bunched tongue shapes than retroflex ones, which is suggestive of a relationship between the tongue and the lips. Although lip protrusion increased in hyperarticulated productions across the board, bunched tongue shapes were still predicted to have significantly more lip protrusion than retroflex ones. We suggest a trading relationship between the degree of lip protrusion and the size of the sublingual cavity. Retroflex tongue shapes with the tongue tip raised create more space underneath the tongue than bunched ones. Bunched /r/ users may compensate for their smaller sublingual space with increased lip protrusion. To attain lower F3 values in hyperarticulation, while retroflex users may increase the size of the sublingual space with increased retroflexion, this strategy is not available to bunched /r/ users and so, they have to make do with increased lip protrusion to lower F3.

3. Is Anglo-English /r/ produced with a labiodental articulation even in the cases where there is an observable tongue body gesture?

Yes. By comparing the lip postures for /r/ and /w/, we concluded that /r/ has a specific labial gesture which differs from that of /w/, even when /r/ is produced with an observable tongue body gesture. While /w/ is produced with horizontal labialisation, /r/ is produced with vertical labialisation. This vertical labialisation allows speakers to protrude their lips without employing close lip rounding, which may enhance the acoustic output of /r/ by maintaining a close proximity between F3 and F2. The lips are extended up and outwards for /r/, which results in an approximation of the inside of the bottom lip with the upper top teeth, which is indicative of a labiodental articulation.

4. Is the labial posture for Anglo-English /r/ perceptually salient?

Yes. The results from the perception experiment indicate that the /r/ typical labial posture defined in Experiment 2 is visually salient. Participants are able to identify /r/ and /w/ from lip reading alone with a very high degree of accuracy. Sensitivity to the /r/-/w/ contrast is not only

enhanced with the addition of visual cues, but the visual cues seem to be more informative than the auditory ones. There was no significant difference between visual-only and audio-visual perception of the /r/-/w/ contrast, which indicates that the phonetic information contained within the visual cues is salient enough on its own to allow observers to accurately distinguish between /r/ and /w/. Furthermore, in incongruous audio-visual trials, the visual cue of the lips for /r/ generally overrides the auditory cue, which suggests that the labial posture for Anglo-English /r/ is unambiguous with respect to /w/ and is highly informative to observers in their perceptual identification of /r/.

## 7.2 THEORETICAL IMPLICATIONS

### *7.2.1 Phonetic accounts of labialisation*

Our review of the literature on the phonetics of labialisation led us to question the appropriateness of the term 'lip rounding' for certain descriptions of speech segments. In some cases, sounds that are typically described as 'rounded', such as front rounded vowels, may not actually be produced with rounded lips. The restrictive nature of the term may have also led to the labiality of certain segments to be overlooked, such as the case of Japanese /u/, which may actually be produced with labialisation, despite what phonological accounts would suggest. We thus proposed the use of *labialisation* as a less restrictive, more phonetically neutral label that would allow for more detailed descriptions of how a reduction in lip area might be implemented. Indeed, we defined labialisation as a secondary labial articulation in consonants and vowels resulting in a reduction of the overall lip area. We gleaned from the literature that there are two main labialisation strategies, both of which may be accompanied by lip protrusion: horizontal labialisation and vertical labialisation. We observed that the two strategies may vary with respect to the size of the lip opening. Horizontal labialisation produces a small, rounded opening, while a larger, slit-like opening is formed using vertical labialisation. Differences in the size of the lip opening and in the degree of lip protrusion may have an impact on the resulting acoustics. For example, Fant's nomograms predict that in the case of a narrow lingual

constriction in the pre-palatal region, decreases in lip area particularly affect the third formant, i.e., in the case of rounded close-front vowels such as [y]. Conversely, it is the second formant that is most affected by decreases in lip area when the lingual constriction is more posterior, i.e., for [u]. These acoustic differences may explain why vertical labialisation generally occurs in front vowels while horizontal labialisation occurs in back vowels. Front vowels are produced with a relatively large lip area to maintain a minimal distance between F2 and F3, while back vowels are produced with a small lip area formed by close lip rounding in order to keep F2 maximally low.

The results from Experiment 2 in which we investigated labialisation in Anglo-English /r/ and /w/ further support our suggestion that the phonetic implementation of 'lip rounding' is not as simple as what the binary phonological feature [± round] would indicate. We suggest that while /w/ requires close rounding in order to produce a maximally low second formant, /r/ is produced with a larger lip area in order to avoid over-lowering F2 and to produce a minimal distance between F2 and F3. /w/ unexpectedly shares a similar labialisation strategy to the one detailed in the literature for the back vowel /u/. Anglo-English /r/, on the other hand, seems to share closer labial properties with front vowels. In the literature on English /r/, it is generally agreed that English /r/ involves 'lip rounding'. In actual fact, our results indicate that /r/ is produced not with rounded lips, but with an approximation of the bottom and top lip initiated by the raising of the bottom lip, at least in Anglo-English. This labial gesture seems to have acoustic consequences.

We therefore suggest that articulatory phonetic studies take a closer look at the lips and investigate how exactly lip rounding or labialisation is implemented. On a methodological level, we have shown that techniques from deep learning allow us to segment the lips from the rest of the face in front images of the speaker with a high degree of accuracy. This technique requires less preparation and post-processing of the data than in other contour detection and extraction techniques, which is an undeniable advantage. Our extraction technique with deep learning also provided automatic measures of lip dimensions. We found that this automatic technique was far less time consuming and more reproducible than taking measures of the lips

by hand. We are currently working on extending this technique to dynamic measurements of the lips as well as static ones. We recommend future studies incorporate deep learning into their analysis of not just the lips, but of all sorts of phonetic data. It is hoped that deep learning will gradually allow the researcher to spend less time on data preparation and processing and more time on interpreting results.

### *7.2.2 A phonetic account of the change in progress towards labiodental* /ɹ/

As we have described throughout this thesis, a change in progress is currently underway towards non-lingual labiodental productions of /ɹ/ in Anglo-English. Although phonetic accounts are lacking, this change in progress is defined as the loss of the lingual articulation of the post-alveolar approximant, leaving the remaining labial articulation to form the primary constriction (Docherty & Foulkes, 2001; Foulkes & Docherty, 2000; Jones, 1972). By comparing the labial postures of /ɹ/ and /w/ in Anglo-English speakers, our results confirmed that the labial posture for /ɹ/ shares similar articulatory properties to a labiodental articulation. This finding therefore supports the claim that labiodental variants may continue to emerge if the labial component is retained at the expense of the lingual one.

However, we argue that a labiodental articulation has developed to accompany lingual articulations of /ɹ/ due to increased exposure to labiodental variants and to the ensuing pressure to maintain a perceptual contrast with /w/. In acoustic terms, what distinguishes post-alveolar articulations of /ɹ/ from /w/ is F3. However, as Dalcher et al. (2008) proposed, non-lingual labiodental variants of /ɹ/ generate high F3 values which do not contrast with the high F3 produced for /w/. As a result, Dalcher et al. argued that the increase in /ɹ/ variability with respect to its third formant may have catalysed a cue-shift from F3 to F2 in the perception of the /ɹ/-/w/ contrast in England. The second formant is lower for [w] than it is for [ʊ]. We proposed that a labiodental-like articulation has developed for lingual articulations of /ɹ/ in order to keep F3 maximally low and F2 maximally high. This labiodental articulation allows speakers to protrude their lips without close rounding. An overly rounded lip posture for /ɹ/ would have a lowering effect on F2, which may cause perceptual uncertainty with /w/. In Englishes that are

unexposed to F3 variation caused by labiodental /r/ such as American English, F3 remains the most salient perceptual cue (Dalcher et al., 2008). We propose that this allows American English productions of /r/ more freedom when it comes to the implementation of labialisation and a recent study which described the labial posture(s) for American English /r/ indeed presents more variability than in our Anglo-English speakers (B. J. Smith et al., 2019). We suggest then that in the pressure to retain a perceptual contrast between /r/ and /w/ caused by exposure to high-F3 labiodental /r/ in Anglo-English, lingual articulations of /r/ may have inadvertently become more labiodental.

Despite having argued that the evolution of an /r/ typical labial gesture was catalysed by pressure to optimise the auditory perception of the /r/-/w/ contrast, results from our perception study indicate that auditory perception of the contrast remains problematic, perhaps due to enhanced exposure to labiodental /r/. Anglo-English listeners have to tolerate such a high degree of acoustic variability for /r/ that low F2, high F3 productions of /w/ may be reconstructed as /r/. However, we suggest that the /r/ specific labial gesture optimises the perception of the /r/-/w/ contrast visually. Perception of the contrast is dominated by visual rather than auditory cues. As a result, we may predict that the change to non-lingual labiodental /r/ will continue in Anglo-English given the relative perceptual certainty of the visual cues of the lips contrary to the acoustic ones. Our results thus support Docherty and Foulkes (2001) hypothesis that labiodentalisation will continue to increase due to 'the heavy visual prominence of the labial gesture' (p. 183).

In our review of the literature on audio-visual speech perception presented in Chapter 1, we observed that previous studies have shown that children and infants are sensitive to visual speech cues. It was argued that visible cues of adult articulations may be utilised by children as visual feedback during the acquisition process (as proposed by Lin & Demuth, 2015). We could therefore imagine a scenario in which children acquiring Anglo-English /r/ are confronted with a prominent visual cue in adult speakers. Knight et al. (2007) presented acoustic data from one speaker acquiring Anglo-English between the ages of 3;8 and 3;11. They found that progress towards adult-like post-alveolar approximant /r/ is manifested through a gradual raising of

F2 and a lowering of F3. They argued that their data suggest that 'developing speakers move gradually away from [w]-like articulations of /r/ to more adult-like articulations, producing a labiodental variant along the way' (p. 1581). Although the authors did not mention the influence of the visual speech cues, visual feedback may show developing speakers that the typical labial posture for /r/ differs from that of /w/. There is therefore a gradual shift from strong [w]-like lip rounding for /r/ to a more labiodental-like posture with the help of visual feedback, followed by the acquisition of the accompanying lingual component in individuals who acquire lingual /r/. We may imagine that this latter step may gradually erode in children acquiring Anglo-English, given the visual salience of the labial cue, not to mention the fact that acquisition of the lingual articulation is undeniably complex, hence why it is often a target in phonological intervention in American English (Adler-Bock et al., 2007).

### *7.2.3 The nature of speech perception*

As we showed in Chapter 1, it is now widely accepted that speech is perceived using information from multiple senses. In this thesis, we have focused on the impact of audition and sight. We know that seeing the speaker's articulatory movements, especially from the lips, substantially improves the perception of speech when the auditory information is degraded, caused either by hearing loss or by background noise. This is because visual speech gestures provide cues to the place and perhaps to the manner of articulation of certain speech sounds by presenting information about the position of the speaker's articulators. The results from our perception study indeed indicate that perceptual sensitivity to the /r/-/w/ contrast is significantly enhanced when participants can see the speaker as opposed to just hearing her productions presented in noise. This research therefore provides further evidence for the multimodal nature of speech perception.

It is generally agreed that auditory speech is more informative than visual speech. Indeed, in English there are fewer visemes than there are phonemes. However, our results indicate that the visual cues from the lips for the /r/-/w/ contrast may actually be more perceptually salient than the auditory ones. This result suggests that in some cases, visual cues may not just be

complementary to the auditory ones. Visual information may provide crucial phonetic cues, which enable listeners to disambiguate similar-sounding speech sounds. Without the visual cues from the lips, productions of /w/ may be incorrectly reconstructed as /ɹ/ by Anglo-English listeners, given the acoustic proximity of [w] to labiodental /ɹ/ ([ʋ]). However, the lip postures for /ɹ/ and /w/ are visually distinct in Anglo-English, allowing listeners to distinguish between the two far more easily from the visual cues than the auditory ones.

We have also observed that productions of /ɹ/ which still have an observable lingual articulation are produced with a lip posture that closely resembles a labiodental articulation. Consequently, the most frequent realisations of prevocalic /ɹ/ in England, i.e., the post-alveolar and the labiodental approximants, may share a very similar labial posture. While these variants have a common lip posture, their resulting acoustic properties differ considerably. As a result, the unifying feature and perhaps the most stable phonetic property of the majority of articulations of Anglo-English /ɹ/ may now be the labial posture. In this case, visual speech cues may thus provide more informative phonetic cues than the more ambiguous and variable auditory ones. A connection may be made between the results presented in this thesis and the ones in Traunmüller and Öhrström (2007) concerning audio-visual perception of the /i/-/y/ contrast in Swedish. They observed that participants relied more heavily on visual cues than on auditory ones for this particular contrast. Traunmüller and Öhrström suggested that the perception of any given feature is dominated by the modality which provides the most reliable information. Our results support this claim. We thus conclude that although a spoken message is usually transmitted from the speaker to the listener via the acoustic signal generated by the speaker's vocal movements, when the auditory cues for any given contrast are ambiguous, listeners may look to alternative phonetic cues from other modalities, such as the ones provided by vision, to better disambiguate the contrast. If the visual modality provides more informative phonetic cues to the contrast than the auditory modality, speech perception may be dominated by the listener's eyes and not by the ears.

### 7.2.4 *The evolution of phonological sound systems: Towards an Audio-Visual Enhancement Hypothesis*

We may ask whether the availability of perceptually salient visual speech cues in the event of auditory ambiguity is simply the result of a fortuitous evolutionary accident, or whether it is the result of a phonological system specifically developed to exploit the multimodal nature of speech perception. In other words, are phonological systems optimised for both auditory and visual speech perception? The conclusions we have drawn from the three experiments presented in this thesis would certainly point in that direction. Indeed, one of our main conclusions is that the labial gesture in Anglo-English /r/ has evolved in order to reinforce the phonological contrast with /w/. The idea that speech segments are enhanced to optimise the perception of phonological contrasts is not new. For example in their *Auditory Enhancement Hypothesis*, Diehl and Kluender (1989) proposed that the phonemic inventories of the world's languages are determined by considerations of maximising perceptual distinctiveness. As Diehl and Kluender noted, it seems likely that phonological inventories have evolved to be 'fairly robust signalling devices' (p. 123). Consequently, there is a tendency for languages to select properties of speech sounds that reinforce phonological contrasts. A typical example involves the use of lip rounding in back vowels. Back vowels are generally produced with lip rounding across the board because rounding auditorily enhances the tongue backing gesture by contributing to F2 lowering. In contrast, relatively fewer instances of lip rounding occur in front vowels in the phonemic inventories of the world's languages because lip rounding counteracts the acoustic effect of tongue fronting (i.e., a high F2). In the case of back vowels, given the fact that lip rounding may in some ways be considered an enhancement mechanism, this lip rounding is arguably somewhat acoustically redundant. A low F2 may still be achieved without lip rounding via tongue backing, although unrounded back vowels would likely result in higher F2 values than their rounded counterparts.

In terms of language change, the Auditory Enhancement Hypothesis would predict new features to arise when an existing phonological contrast is insufficiently perceptually salient. Parallels may thus be drawn between this framework and our conclusion that the labial gesture

in Anglo-English /ɹ/ has evolved in order to reinforce the phonological contrast with /w/. However, we point out that the Auditory Enhancement Hypothesis is, by definition, concerned exclusively with auditory speech perception. The possible effect of visual cues has yet to be accounted for, although a logical extension which we put forward here would be an *Audio-Visual Enhancement Hypothesis*. If an auditory contrast is insufficiently salient (as in Anglo-English /ɹ/-/w/), phonetic cues may be enhanced both auditorily and/or visually. Phonological systems may thus evolve to exploit the audio-visual nature of speech perception.

Other evidence for the optimisation of phonological systems according to visually salient phonetic features may be drawn from commonly occurring phonological contrasts in phonemic inventories. As we observed in Chapter 1, almost all of the world's languages contrast bilabial and coronal stops and yet the articulation of these contrasts tends to produce acoustically similar sounds, e.g., [m] versus [n]. However, the visual distinction between bilabial and coronal articulations may maximise the perceptual distinctiveness of these sounds, which may explain why they occur so frequently, despite the limitations their similar auditory cues may cause.

A link may also be made between our Audio-Visual Enhancement Hypothesis and the results presented in Havenhill (2018) and Havenhill and Do (2018). They examined the audio-visual perception of the COT-CAUGHT contrast in American English, which is currently undergoing a merger in some dialects. It was found that despite having a similar acoustic output, productions of /ɔ/ with visible rounding were more perceptually salient than those without rounding. In a similar conclusion to the one we present in this thesis, Havenhill (2018) and Havenhill and Do (2018) argued that visual cues may play a role in the shaping of phonological systems by inhibiting misperception of the speech signal in cases where two sounds are acoustically similar. In this instance, a visible labial cue is retained despite the apparent merging of the phonological contrast. They proposed that phonological systems may be 'optimised' to enhance both auditory and visual perceptibility.

Another pertinent diachronic sound change in English may involve /u/-fronting, which is reportedly observed in Englishes worldwide. As described in Section 3.6, in terms of acoustics, /u/-fronting manifests itself as the raising of the second formant. As the term *fronting* implies, it

is generally assumed that this /u/-fronting is the result of the fronting of the palatal constriction from an originally back position. However, a similar acoustic effect of F2 raising may also be a consequence of lip unrounding. Harrington et al. (2011) assessed the lingual and labial articulation of /u/ in SSBE speakers and found that fronting indeed affects the position of the tongue, and not the rounding of the lips (although see Lawson et al., 2019, for decreased lip protrusion in fronted /u/ in Scottish English). We suggest that given the optimising effect of visual cues on the perception of phonological contrasts, the labial gesture may have a privileged status, meaning that it is retained in diachronic sound change. The results presented in Havenhill (2018) and Havenhill and Do (2018) appear to support this account.

Finally, this 'privileged' status of visible articulations may have origins in the way in which spoken language first came about. Some theorists suggest that the first languages were gestural as opposed to vocal in nature. The persisting contribution of visual facial cues in speech perception today and potentially in the evolution of phonological systems may be a remnant of these original gestural languages, which were likely perceived entirely visually. As such, we contend that the maximisation of phonological contrasts via visual cues is no evolutionary accident and is the consequence of the primitive nature of audio-visual speech perception, as proposed by Rosenblum (2008a). Speech has evolved and continues to evolve to be both heard and seen, and the evolution of a generalised labial gesture in Anglo-English /r/ is an example of a change in progress which exploits the multimodal nature of speech perception in order to maximise a phonological contrast.

## 7.3 CONTRIBUTIONS

This thesis has first and foremost begun to fill the distinct gap in the literature on the phonetics of Anglo-English /r/. We have supplied acoustic, articulatory and perceptual evidence to show that despite sharing many similar phonetic characteristics with rhotic varieties such as American English and Scottish English, Anglo-English /r/ is in some ways unique and warrants our attention. Articulatory data provided by Ultrasound Tongue Imaging support the hypothesis that although non-rhotic Englishes may produce /r/ with a multitude of tongue

shapes, higher rates of retroflexion occur in non-rhotic Englishes than in rhotic varieties, as proposed by Heyne et al. (2018). The major take home message from this thesis is that the labial articulation of Anglo-English /r/ plays a pivotal role in both its production and perception. We have observed a relationship between tongue shape and the degree of lip protrusion, which we equate to the size of the front cavity, which is a novel finding for English /r/. Lip protrusion contributes to the acoustics of the post-alveolar approximant by playing a part in the lowering of the third formant. Speakers make active adaptations to their speech patterns for /r/ in order to enhance its perceptibility including increased lip protrusion and retroflexion. We have also described the evolution of a specific labial posture for Anglo-English /r/ which allows speakers to optimise their production for enhanced auditory and visual perception of the /r/-/w/ contrast. These findings have theoretical implications for phonetic descriptions of lip rounding, and for the role of visual speech cues in speech perception generally as well as in diachronic sound change and in the evolution of phonological sound systems. Finally, on a methodological level, we have shown that techniques from deep learning may be applied to phonetic data to produce interpretable and meaningful analyses, which is a promising research avenue for the future.

## 7.4 Limitations and future directions

As described at the end of Chapter 4 (Section 4.4.2, p. 163), although the results from Experiment 1 point towards a possible articulatory compensation strategy involving increased lip protrusion to extend the front cavity for /r/, more articulatory data, ideally from a more robust imaging technique that could provide vocal tract dimensions e.g., real-time MRI, is required to further confirm our claim. Indeed, a limitation to our study is the fact that the sublingual space is not measurable from ultrasound data. Furthermore, there may well be a three-way trading relation between the size of the sublingual space, palatal constriction location, and degree of lip protrusion, all of which would extend the front cavity, which could be researched in the future. Furthermore, we see no reason why the use of lip protrusion as a compensation strategy for /r/ could not be extended to other syllabic contexts and to other varieties of English, which

could also be the object of further study.

The experiments presented in this thesis present a synchronic account of the production and perception of Anglo-English /r/. Although the speakers recruited for the production experiments had a relatively wide range of ages (from 18 to 55), the data was not stratified enough for age to justify considering it as a potential predictor of tongue shape or lip protrusion. The data point towards a possible effect of age in that the most variable tongue shapes tended to occur in the youngest speakers. However, that does not mean to say that all the older speakers used one tongue shape exclusively. In a study presented at the most recent 6$^{th}$ edition of the *R-atics Colloquium* in Paris, the international conference dedicated to the study of 'r'-sounds, Strycharczuk, Lloyd, and Scobbie (2019) collected ultrasound data at a public outreach event from 36 SSBE speakers aged between 16 and 78 for /r/ and observed a significant effect of age. Younger speakers were more likely to produce tip-down tongue shapes than older ones. Future studies could continue to investigate how the lingual articulation of /r/ may have changed over the years by considering a larger cohort of speakers stratified for age. If tip down shapes are a recent innovation, we may predict more labiality in younger than in older speakers.

The perception experiment indicated that participants are more sensitive to the /r/-/w/ contrast in the visual-only modality than the auditory-only one. We have interpreted this finding to indicate that the visual cues may be more phonetically informative than the auditory ones. We must stress however that the auditory cues were masked in noise, which would naturally make auditory perception more challenging than in optimal listening conditions. To further enhance this claim, it might be worth running the study again without the addition of background noise. We predict that auditory sensitivity would increase, although /w/ productions may still be reconstructed as /r/ but perhaps to a lesser degree than in noise. The rate of visual capture may also lower without noise because decreasing the intensity of auditory cues or masking them with noise has been found to increase incidences of the McGurk Effect (Colin et al., 2002; Fixmer & Hawkins, 1998; Sekiyama et al., 2003). Despite the limitations associated with the addition of noise, 96% of all the visual-only responses for the /r/-/w/ contrast were correctly identified, which shows that listeners are highly sensitive to visual cues.

However, the stimuli for the perception experiment were produced by one speaker. In future studies, it may be worth including perception stimuli from multiple talkers to ensure that the visual effect for /ɹ/-/w/ is robust, despite potential inter-speaker variation. One could also consider the effect of visibility of the visual cues on the perception of /ɹ/ and /w/. This could be achieved by presenting listeners with images of the speaker taken at different distances, or by varying the quality of the images under presentation.

We have argued that a specific labial posture for /ɹ/ has evolved in Anglo-English due to high exposure to non-lingual labiodental variants and that this labial posture has perceptual consequences in Anglo-English listeners. We propose that a generalised labial posture for /ɹ/ may not occur in Englishes where labiodentalisation is not common, such as American English. This is because auditory perception of the /ɹ/-/w/ contrast should be relatively less ambiguous, meaning that the contrast does not need enhancing with the lips. These claims require further investigation. Although B. J. Smith et al. (2019) showed that the labial posture for American English indeed varies across speakers, we do not know how American English listeners fare when it comes to perception. We plan to consider the audio-visual perception of the /ɹ/-/w/ contrast in American English in the future. We would predict that sensitivity to the visual cues for the /ɹ/-/w/ contrast is less high in American English than in Anglo-English because the labial postures for /ɹ/ and /w/ may be more ambiguous. Similarly, we would expect fewer cases of visual capture to occur in American English because perception is likely dominated by the auditory as opposed to the visual phonetic cues. We would also predict American English listeners to perceive labiodental variants of /ɹ/ as /w/, given their lack of phonetic experience of labiodental /ɹ/ and the acoustic similarity between [w] and [ʋ]. In addition, the /ɹ/ bias we observed in the auditory perception of the /ɹ/-/w/ contrast in Anglo-English would likely not occur in American English listeners, given their lack of experience of labiodental /ɹ/, which we propose is the reason for said bias in Anglo-English.

Finally, future studies may consider the role of the visual cue of the lips in the acquisition of /ɹ/ in speakers learning English as a second language. For example, it has been observed that French listeners present perceptual difficulties with American English /ɹ/, which tends to be

assimilated to /w/ (Hallé, Best, & Levitt, 1999). From a phonological standpoint, French learners should not exhibit such a problem because French has two equivalent phonemes. However, the French [ʁ] is dissimilar to English /r/ both in acoustic and articulatory terms. Hallé et al. argued that the labial gesture present in American English /r/ may show more similarity with French [w] than [ʁ], leading American English /r/ to be perceived as /w/-like. Interestingly, in the same study, French listeners' discrimination of American English /w/-/j/ was significantly better than native speakers'. /w/ and /j/ contrast in French and have almost identical phonetic realisations to American English /w/-/j/, as Hallé et al. (1999) pointed out. They speculated that increased sensitivity to the /w/-/j/ contrast in French may stem from greater sensitivity to semi-vowels more generally, given the richer phonological system in French, which includes another semi-vowel /ɥ/. However, Bohn and Best (2012) proposed another possible systemic factor: a difference in vowel systems. Just like in the French participants in Hallé et al. (1999), Bohn and Best (2012) found that the discrimination of American English /w/-/j/ was better in German and in Danish listeners than in native English speakers. While German and Danish have fewer semi-vowels than French, all three languages have front-rounded vowels in their phonological inventories that English lacks. Bohn and Best (2012) therefore concluded that

> the highly overlearned sensitivity to lip rounding distinctions in vowels enables native listeners of languages with such distinctions to discriminate an approximant contrast at near ceiling, if this approximant contrast is importantly differentiated through lip rounding, as is /w/-/j/ but not, e.g., American English /w/-/r/. (p. 19)

We argue that contrary to American English, labialisation is implemented in two distinct ways for Anglo-English /r/ and /w/. While /w/ is produced with the labial posture widely-associated with back-rounded vowels, /r/ is produced with a posture which accompanies front-rounded vowels. We therefore predict that native speakers of languages like French, German and Danish may have heightened sensitivity to the visual cues of /r/ and /w/ in Anglo-English, contrary to native speakers of languages like Japanese, which do not have phonological rounding. Future studies could not only test this claim with audio-visual perception experiments, but could consider whether explicit phonetic training which highlights the difference in labialisation

between /r/ and /w/ may improve the perception and production of /r/ in learners of Anglo-English. Indeed, the pronunciation of /r/ poses a challenge to many learners of English and as Scobbie (2006) remarked, may create the impression of a strong foreign accent when produced incorrectly.

## 7.5 Conclusion

If we revisit the citation from Docherty and Foulkes (2001) which provides the epigraph of this thesis, our results confirm the importance of attending to labialisation in phonetic descriptions of English /r/. Not only do the lips play a role in enhancing the auditory effect of rhoticity, but they also contribute to optimising the perception of /r/ visually. We have suggested that /r/ is produced with a specific labial posture which may be unique to Anglo-English. Exposure to labiodental articulations of /r/ which lack a lingual constriction has resulted in perceptual ambiguity between /r/ and /w/ in England. Listeners must tolerate such a high degree of acoustic variation for /r/ that productions of [w] may be reconstructed as /r/. We propose that the lips enhance the visual saliency of Anglo-English, which may help maintain the phonological contrast between /r/ and /w/. While auditory perception of the /r/-/w/ contrast may pose a challenge to English listeners, prominent visual cues from the speaker's lips allow them to disambiguate the contrast with an exceptionally high degree of accuracy. In proposing an Audio-Visual Enhancement Hypothesis, we contend that languages select audio-visual properties of speech sounds to reinforce phonological contrasts. Phonological systems may thus evolve to exploit the primitive multimodal nature of speech perception: speech has evolved and continues to evolve to be both heard and seen.

# Appendices

# Production experiments

## Contents

Queen Margaret University
EDINBURGH

université PARIS DIDEROT PARIS 7


***Speech recognition using tongue and lip movement during speech***

## Consent Form

---

I have read and understood the information sheet and this consent form. I have had an opportunity to ask questions about the project.

I understand that I am under no obligation to take part in this study, and that I have the right to withdraw from this study at any stage before or during data collection, without giving any reason.

**Please indicate that you give consent to take part in this study by ticking the YES box.**

I agree to participate in this study and that audio recordings of my voice, ultrasound recordings of my tongue and video of my lip movements can be stored indefinitely and used for academic purposes (e.g. analysis, research, academic conference presentations, public engagement lectures, publications and future applications for research funding) — Yes ❒ No ❒

**Please indicate whether you give consent to anonymised audio recordings, ultrasound tongue image recordings and lip video created during this study to be used in any of the following ways.**

They can be used in teaching at *Queen Margaret University* (QMU) and the *University of Paris Diderot* (UPD). — Yes ❒ No ❒

They can be copied for analysis by other researchers outside QMU/UPD for their own academic research projects with permission of the current research team. — Yes ❒ No ❒

They can be broadcast to an audience on laboratory open days, science festivals and other public, non-professional talks and presentations. — Yes ❒ No ❒

Selected recordings can be made publicly available on the internet. — Yes ❒ No ❒

Name of participant

_____

Signature

_____

Investigator

_____

Date: …../……/……

Further information is available from: Hannah King hannah.king@univ-paris-diderot.fr

One copy to be retained by the researcher, one copy to be kept by the participant.

Participant name: _____

Participant identifier: _____

Age:_____

Gender: M / F delete as appropriate


Please indicate with an asterisk ∗ on the map, the location where you have lived longest.

Please write the name of the place where you have lived longest here:

_____

Please add crosses **x** indicating any locations where you have lived for more than a year, and write the locations below:

_____
_____
_____
_____
_____
_____

Please underline your level of education:

primary school

secondary school

further education (college)

higher education (university)

postgraduate degree

Please list any other languages you speak (apart from English) and your proficiency in each (beginner, intermediate, upper intermediate, advanced, mother tongue)

_____
_____
_____
_____
_____
_____

# Perception experiment

B

## Contents

## B.1 FILLERS AND CONTROL STIMULI

| /th/ | /s/ | /h/ |
|---|---|---|
| *thee* | *he* | *see* |
| *this* | *his* | *siss* |
| *thick* | *hick* | *sick* |
| *that* | *hat* | *sat* |
| *thack* | *hack* | *sack* |
| *thank* | *hank* | *sank* |
| *they* | *hay* | *say* |
| *thigh* | *high* | *sigh* |
| *thaw* | *hoar* | *saw* |
| *thawed* | *hoard* | *sword* |
| *thumb* | *hum* | *sum* |
| *though* | *hoe* | *so* |

**Table B.1:** *Filler and control words comprising 36 monosyllabic minimal pairs contrasting /th/, /s/ and /h/ word-initially.*

## B.2 STIMULI PER GROUP

Participants were presented with one of two word lists in the perception experiment. The following tables present the test words for the two groups according to phonological contrast and modality.

| | Lexical set | | | | | |
|---|---|---|---|---|---|---|
| Contrast | FLEECE | KIT | DRESS | TRAP | PRICE | FACE |
| /r-w/ | *reed* | *wit* | *red* | *wag* | *rise* | *wait* |
| /r-l/ | *reek* | *lick* | *rent* | *lack* | *right* | *lake* |
| /w-l/ | *leak* | *wick* | *lent* | *whack* | *light* | *wake* |

**Table B.2:** *Test words presented in the auditory-only modality for Group 1 and in the visual-only modality for Group 2*

| | Lexical set | | | | | |
|---|---|---|---|---|---|---|
| Contrast | FLEECE | KIT | DRESS | TRAP | PRICE | FACE |
| /r-w/ | *week* | *rick* | *went* | *rack* | *white* | *rake* |
| /r-l/ | *lead* | *rit* | *led* | *rag* | *lies* | *rate* |
| /w-l/ | *weed* | *lit* | *wed* | *lag* | *wise* | *late* |

**Table B.3:** *Test words presented in the auditory-only modality for Group 2 and in the visual-only modality for Group 1*

| | Lexical set | | | | | |
|---|---|---|---|---|---|---|
| Contrast | FLEECE | KIT | DRESS | TRAP | PRICE | FACE |
| /r-w/ | *weed* | *rit* | *wed* | *rag* | *wise* | *rate* |
| /r-l/ | *reed* | *lit* | *red* | *lag* | *rise* | *late* |
| /w-l/ | *lead* | *wit* | *led* | *wag* | *lies* | *wait* |

**Table B.4:** *Test words presented in the congruous audio-visual modality for Group 1.*

| Contrast | Lexical set | | | | | |
|---|---|---|---|---|---|---|
| | FLEECE | KIT | DRESS | TRAP | PRICE | FACE |
| /r-w/ | *reek* | *wick* | *rent* | *whack* | *right* | *wake* |
| /r-l/ | *leek* | *rick* | *lent* | *rack* | *light* | *rake* |
| /w-l/ | *week* | *lick* | *went* | *lack* | *white* | *lake* |

**Table B.5:** *Test words presented in the congruous audio-visual modality for Group 2.*

| Lexical set | Auditory cue | Visual cue |
|---|---|---|
| FLEECE | *reed* | *weed* |
| | *weed* | *reed* |
| | *reek* | *week* |
| | *week* | *reek* |
| KIT | *rit* | *wit* |
| | *wit* | *rit* |
| | *rick* | *wick* |
| | *wick* | *rick* |
| DRESS | *red* | *wed* |
| | *wed* | *red* |
| | *rent* | *went* |
| | *went* | *rent* |
| TRAP | *rack* | *whack* |
| | *whack* | *rack* |
| | *rag* | *wag* |
| | *wag* | *rag* |
| PRICE | *right* | *white* |
| | *white* | *right* |
| | *rise* | *wise* |
| | *wise* | *rise* |
| FACE | *rate* | *wait* |
| | *wait* | *rate* |
| | *rake* | *wake* |
| | *wake* | *rake* |

**Table B.6:** *Test words presented in the incongruous audio-visual modality for both groups (Groups 1 and 2).*

```
# getDurationsExtend.praat

# Hannah King

# This script opens all sound files contained within one folder,
double checks that each sound has an associated text grid. If
there is no text grid, Praat creates one with one interval tier
and asks the user to segment each word. A table is created which
contains the duration of the entire sound file, as well as the
duration of the word and of the intervals of silence preceding
and following each word. The script then finds the longest
silence intervals before and after each word in all the files
and extends each sound file and associated text grid to these
maximum lengths. The extended sound files and text grids are
saved in a new folder, along with the final table containing all
extracted duration values.

# path to folder containing raw sound files
path$ = "C:\Desktop\Perception\Stimuli\"

# create a table
table = Create Table with column names: "duration", 0, "fileName
wordStart wordLength preSilence postSilence addPre addPost
orignalfileLength"

# create list of all sound files
soundFiles = Create Strings as file list: "soundFiles", path$ +
"\" + "*.wav"

# get number of sound files
numberFiles = Get number of strings

# open sound files in folder
for allFiles from 1 to numberFiles
    selectObject: soundFiles
    soundName$ = Get string: allFiles
    sound = Read from file: path$ + "\" + soundName$

    # add file name to table
    selectObject(table)
    Append row
    current_row = Get number of rows
    Set string value: current_row, "fileName", soundName$ - ".wav"

    # check textgrid exists. If it doesn't, creates one with one
    tier and asks the user to segment at the word level.
    Textgrid is then saved.
    textgridName$ = (path$ + soundName$ - ".wav") + ".TextGrid"
        if not fileReadable (textgridName$)
        selectObject: sound
        tg = To TextGrid: "word", ""
        selectObject: sound, tg
        View & Edit
            editor: tg
                pauseScript: "Segment word. Click continue."
                Save TextGrid as text file: (path$ + soundName$
                - ".wav") + ".TextGrid"
```

```
                Close
            endeditor
        else
            tg = Read from file: (path$ + soundName$ - ".wav") +
            ".TextGrid"
        endif

    # find length of entire sound file
    selectObject: tg
    orignalfileLength = Get total duration

    # Get number of intervals in first tier of textgrid (i.e.,
    called 'word')
    selectObject: tg
    numberWordIntervals = Get number of intervals: 1

    # find duration of all three intervals in tier 1 ('word')
    for 1 to numberWordIntervals
        selectObject: tg

        # pre-word silence interval (first interval on tier 1)
        preStart = Get start time of interval: 1, 1
        preEnd = Get end time of interval: 1, 1
        preSilence = preEnd - preStart

        # word interval (second interval on tier 1)
        wordStart = Get start time of interval: 1, 2
        wordEnd = Get end time of interval: 1, 2
        wordLength = wordEnd - wordStart

        # post-word silence interval (third interval on tier 1)
        postStart = Get start time of interval: 1, 3
        postEnd = Get end time of interval: 1, 3
        postSilence = postEnd - postStart

        # add values to table
        selectObject(table)
        current_row = Get number of rows
        Set string value: current_row, "fileName", soundName$ -
        ".wav"
        Set numeric value: current_row, "orignalfileLength",
        orignalfileLength
        Set numeric value: current_row, "preSilence", preSilence
        Set numeric value: current_row, "wordStart", wordStart
        Set numeric value: current_row, "wordLength", wordLength
        Set numeric value: current_row, "postSilence", postSilence
    endfor

    # remove objects
    selectObject: sound
    plusObject: tg
    Remove
endfor

# now we need to find the maximum length of preSilence and
postSilence from the table
selectObject(table)
```

```
# find max pre-word silence interval
maxPreSilence = Get maximum: "preSilence"
# add 1 ms so that all files get extended
maxPreSilence1 = maxPreSilence + 0.001
maxPreSilence = number(fixed$(maxPreSilence1, 3))
# add to info line
writeInfoLine: "Normalised post word silence: ", maxPreSilence,
" seconds"

# find max post-word silence interval
maxPostSilence = Get maximum: "postSilence"
# add 1 ms so that all files get extended
maxPostSilence1 = maxPostSilence + 0.001
maxPostSilence = number(fixed$(maxPostSilence1, 3))
# add to info line
appendInfoLine: "Normalised post word silence: ",
maxPostSilence, " seconds"

# now we need to calculate the difference between the length of
the original pauses and the normalised ones for each file.
for allFiles from 1 to numberFiles
    selectObject: soundFiles
    soundName$ = Get string: allFiles
    sound = Read from file: path$ + "\" + soundName$
    tg = Read from file: (path$ + soundName$ - ".wav") +
    ".TextGrid"

    # get number of intervals in first tier of textgrid (i.e.,
    called 'word')
    selectObject: tg
    numberWordIntervals = Get number of intervals: 1

    # find interval durations
    for 1 to numberWordIntervals
        selectObject: tg

        # pre-word silence interval (first interval on tier 1)
        preStart = Get start time of interval: 1, 1
        preEnd = Get end time of interval: 1, 1
        preSilence = preEnd - preStart
        addPre = maxPreSilence - preSilence

        # post-word silence interval (third interval on tier 1)
        postStart = Get start time of interval: 1, 3
        postEnd = Get end time of interval: 1, 3
        postSilence = postEnd - postStart
        addPost = maxPostSilence - postSilence

        # add values to table
        selectObject(table)
        current_row = allFiles
        Set numeric value: current_row, "addPre", addPre
        Set numeric value: current_row, "addPost", addPost
    endfor

    # remove objects
```

```
    selectObject: sound
    plusObject: tg
    Remove
endfor

# save table
selectObject: table
Save as tab-separated file: path$ + "rawDurations.txt"

# remove file list
selectObject: soundFiles
Remove

# now we have a table with the relevant durations which we can
use to extend the wav files. We'll make a new path to the folder
where we would like to save the extended wav files
extended$ = "C:\Desktop\Perception\Stimuli\Extended\"

# we'll add another column to the table to include the final
extended length of each file
selectObject: table
numColumns = Get number of columns
lastColumn = numColumns + 1
Insert column: lastColumn, "finalLength"

# open sound and tg files based on table
selectObject: table
number_files = Get number of rows
for allfiles from 1 to number_files
    selectObject: table

    # get the name of the file from the fileName column of the
    table
    filename$ = Get value: allfiles, "fileName"

    # open sound and textgrid
    sound = Read from file: path$ + filename$ + ".wav"
    tg = Read from file: (path$ + filename$ - ".wav") +
    ".TextGrid"

    # we need to know the sampling frequency of the sound file
    to create silence
    selectObject: sound
    samplingFrequency = Get sampling frequency

    # now get the length we need to add at the start of the
    recording from the table
    selectObject: table
    addPre = Get value: allfiles, "addPre"

    # make silence
    myPreSilence = Create Sound from formula: "silence", 1, 0,
    addPre, samplingFrequency, "0"

    # Praat concatenates sounds based on the order in which they
    appear in the list of objects, so we need to make a new
    sound file before we can combine the sound with the silence
```

```
selectObject: sound
sound2 = Copy: "copy"

# select sounds and combine
selectObject: myPreSilence
plusObject: sound2
soundLongPre = Concatenate

#remove original sound, the copy and silence objects
selectObject: myPreSilence
plusObject: sound
plusObject: sound2
Remove

# extend textgrid
selectObject: tg
Extend time: addPre, "Start"

# Praat adds a boundary where the textgrid originally
started. Let's remove it.
selectObject: tg
Remove left boundary: 1, 2

# now let's add silence at the end of the files based on the
values from the table
selectObject: table
addPost = Get value: allfiles, "addPost"

# make silence
myPostSilence = Create Sound from formula: "silence", 1, 0,
addPost, samplingFrequency, "0"

# select sounds and combine
selectObject: soundLongPre
plusObject: myPostSilence
soundLong = Concatenate

# remove silence file and old sound file
selectObject: myPostSilence
plusObject: soundLongPre
Remove

# now let's extend the textgrid
selectObject: tg
Extend time: addPost, "End"

# Praat adds a boundary where the textgrid originally
started. Let's remove it.
Remove right boundary: 1, 3

# scale the times of the new sound and textgrid
selectObject: soundLong
plusObject: tg
Scale times

# get duration of final sound file
selectObject: soundLong
```

```
    durFinal = Get total duration

    # add final duration value to the table
    selectObject(table)
    Set numeric value: allfiles, "finalLength", durFinal

    # save final sound file
    selectObject: soundLong
    Save as WAV file: extended$ + filename$ + ".wav"

    # save textgrid
    selectObject: tg
    Save as text file: (extended$ + filename$) + ".TextGrid"

    # remove sound and tg files
    selectObject: tg
    plusObject: soundLong
    Remove
endfor

# save table as txt file.
selectObject: table
Save as tab-separated file: extended$ + "extendedDurations.txt"

# tell user the script has finished running
appendInfoLine: "All done!"
```

université PARIS DIDEROT PARIS 7 Université de Paris

UNIVERSITY of York

**DEPARTMENT OF LANGUAGE AND LINGUISTIC SCIENCE**
Heslington, York, YO10 5DD, UK
hannah.king@univ-paris-diderot.fr

***Decoding Speech in Noisy Conditions***

Lead researcher: Hannah King, University of Paris – Paris Diderot

## Consent Form

---

This form is for you to state whether or not you agree to take part in the study. Please read and answer every question. If there is anything you do not understand, or if you want more information, please ask the researcher.

Have you read and understood the information leaflet about the study? Yes ❒ No ❒

Have you had an opportunity to ask questions about the study and have these been answered satisfactorily? Yes ❒ No ❒

Do you understand that the information you provide will be held in confidence by the research team, and your name or identifying information about you will not be mentioned in any publication? Yes ❒ No ❒

Do you understand that you may withdraw from the study at any time before the end of the data collection session without giving any reason, and that in such a case all your data will be destroyed? Yes ❒ No ❒

Do you understand that the information you provide may be kept after the duration of the current project, to be used in future research on language? Yes ❒ No ❒

Do you agree to take part in the study? Yes ❒ No ❒

Your name (in BLOCK letters):

_____

Your signature:

_____

Researcher's name:

_____

Date:

_____

One copy to be retained by the researcher, one copy to be kept by the participant.

Participant identifier: _____

Age: _____

Sex:

☐ Female
☐ Male

Are you:

☐ Right-handed
☐ Left-handed

**Origins**

Place of birth
*(i.e. village/town, county, country)*

_____

Where did you spend the most time growing up (i.e. until you were 18 years old)?
*(i.e. village/town, county, country)*

_____

Have you ever lived in another English-speaking country for more than one year? If yes where and for how long?

☐ No
☐ Yes *place:* _____
*duration:* _____
*place:* _____
*duration:* _____
*place:* _____
*duration:* _____

**Education**

What is your level of education (inclusive of qualifications currently in preparation)

☐ Primary school
☐ Secondary school
☐ Further education (6th form/college)
☐ Undergraduate degree
☐ Postgraduate degree

**Languages**

Native language

_____

What language do you speak at home?

_____

Please list all the languages you speak apart from English in order of dominance. Please include your proficiency in each language. Languages spoken at a level lower than intermediate do not need to be included.

| Language | Intermediate (B1) | Upper Intermediate (B2) | Advanced (C1) | Fluent (C2) |
|---|---|---|---|---|
| 1) | ☐ | ☐ | ☐ | ☐ |
| 2) | ☐ | ☐ | ☐ | ☐ |
| 3) | ☐ | ☐ | ☐ | ☐ |
| 4) | ☐ | ☐ | ☐ | ☐ |
| 5) | ☐ | ☐ | ☐ | ☐ |

**Linguistics training**

Do you have any formal (i.e. university level) training in linguistics and/or phonetics? If yes, please give a few details of what the training entailed:

☐ No

☐ Yes

_____

_____

_____

**Speech and hearing**

Have you ever had:

☐ A hearing impairment

☐ A language disorder

☐ A learning disorder

☐ An uncorrected sight problem

If yes, please detail:

_____

Please tick the most appropriate response to the following questions:

| | always | often | sometimes | rarely | never |
|---|---|---|---|---|---|
| Do you feel like you have any hearing problems, which are not currently known or treated? | ☐ | ☐ | ☐ | ☐ | ☐ |
| Do you sometimes find it challenging to have a conversation in quiet surroundings? | ☐ | ☐ | ☐ | ☐ | ☐ |
| Do you find it difficult to understand speech on TV and radio? | ☐ | ☐ | ☐ | ☐ | ☐ |
| Do you find it difficult to follow conversations at dinner parties? | ☐ | ☐ | ☐ | ☐ | ☐ |
| Do you find yourself having to ask people to repeat themselves? | ☐ | ☐ | ☐ | ☐ | ☐ |
| Do you find it hard to have a conversation on the phone? | ☐ | ☐ | ☐ | ☐ | ☐ |

## B.6 Instructions for perception task

Place your hand on the mouse. In this task, you will be presented with some English words, which have been masked with noise. For each word, you will be asked which word you understood from two options. In some cases, you will see a video of the person speaking, in others you will just see an image of her face. There will also be times where you will be presented with the video of the speaker but you will not be able to hear her. It is therefore important that you watch the screen throughout the whole experiment.

Before each word, a cross will appear on the screen, which you should look at, like this...

< example fixation cross >

After the cross, you will be presented with a word automatically. You will then select the word you understood by clicking on one of the two words written on the screen. For example ,if you are presented with...

< example stimulus >

...the following options may be given:

< example responses >

Try to answer as quickly and as accurately as possible. Your first mouse click will be recorded. If you don't answer in time, the programme will automatically advance to the next word. If you are not sure, respond with your best guess. The same word may appear more than once and some words may be more familiar to you than others.

In some cases, the lips will be painted in a bright colour. In these instances, you should respond with the colour of the lips and NOT the word you understood. For example...

< example catch trial >

...The appropriate response was ‘purple’. We will start with a practice round.

< Practice items 1-10 >

We will now begin the experiment for real. If you have any further questions, please speak to the researcher. You will be given opportunities to take a break. Click the mouse to begin.

Université de Paris

# APPENDIX C. LIST OF COPYRIGHTED ITEMS

List of items removed from the full version of the thesis for copyright reasons

## Illustrations, figures, images...

| Caption | Figure N° | Page(s) in manuscript |
|---|---|---|
| Denes and Pinson's (1993) Speech Chain depicting the progression of a speech message from the brain of the speaker to the brain of the listener through the sound waves generated by the speaker's vocal movements. | Figure 1.1 | 8 |
| The geographical distribution of rhoticity based on data from the Survey of English Dialects from the 1950s (left) (Orton & Dieth, 1962) and the English Dialects App from 2016 (right) (from Leemann et al., 2018, p. 12). | Figure 2.1 | 39 |
| Delattre and Freeman (1968)'s taxonomy of tongue shapes for American English and Anglo-English /r/ (from Mielke et al., 2016, p. 103). | Figure 2.2 | 42 |
| Typical examples of tongue configurations for postvocalic /r/ in Scottish English divided into four categories (from Lawson et al., 2013, p. 200). | Figure 2.3 | 49 |
| Locations of nodes and antinodes in a tube open at one end in the unconstricted vocal tract. Perturbation Theory predicts that a constriction at the location of an anitnode (labelled A) in the vocal tract would lower the frequency of the corresponding resonances. Nodes are indicated by the intersections of the sine waves (adapted from Johnson, 2012, Figure 6.7). | Figure 2.4 | 66 |
| Schematisation of possible lip settings according to Laver (1980, p. 37). All lip settings may be accompanied by lip protrusion. The outline of the neutral lip setting is indicated by a dashed line. H – Horizontal; V – Vertical;<br>E – Expansion; C – Constriction | Figure 3.2 | 82 |
| Nomograms from Fant (1989, p. 80) for incremental values of lingual constriction location from the glottis to the lips with a constriction fixed at a narrow area of 0.65 cm2. Curves 1-5 correspond to different lip areas from 8.00 cm2 (no rounding) to 0.16 cm2 (strong rounding). The points of formant merging are circled for [i], [y] and [u]. | Figure 3.3 | 86 |
| Schematised profile views of two labiodental articulations and rounding for [w] (adapted from Catford, 1977, Figure 39). | Figure 5.1 | 173 |

# Bibliography


Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Adachi, T., Akahane-Yamada, R., & Ueda, K. (2006). Intelligibility of English phonemes in noise for native and non-native listeners. *Acoustical Science and Technology*, *27*(5), 285–289. doi: 10.1250/ast.27.285

Adler-Bock, M., Bernhardt, B. M., Gick, B., & Bacsfalvi, P. (2007). The use of ultrasound in remediation of North American English /ɹ/ in 2 adolescents. *American Journal of Speech-Language Pathology*, *16*(2), 128–139. doi: 10.1044/1058-0360(2007/017)

Aloufy, S., Lapidot, M., & Myslobodsky, M. (1996). Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers. *Brain and Language*, *53*(1), 51–57. doi: 10.1006/brln.1996.0036

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research*, *31*(1-2), 111–144. doi: 10.1163/22134808-00002565

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *The Journal of the Acoustical Society of America*, *101*(2), 1078–1089. doi: 10.1121/1.417972

Armstrong, N., & Pooley, T. (2013). Levelling, resistance and divergence in the pronunciation of

English and French. *Language Sciences*, *39*, 141–150. doi: 10.1016/j.langsci.2013.02.018

Articulate Instruments Ltd. (2008). *Ultrasound stabilisation headset users' manual, revision 1.4.* Edinburgh, UK.

Articulate Instruments Ltd. (2014). *Articulate Assistant Advanced ultrasound module user manual, revision 2.16.* Edinburgh, UK.

Ashton, H., & Shepherd, S. (2012). *Work on your accent.* London, UK: Collins.

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1157–1165. doi: 10.1044/1092-4388(2007/080)

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. doi: 10.1177/00238309040470010201

Badin, P., Sawallis, T., & Lamalle, L. (2014). Comparaison des stratégies articulatoires d'un locuteur bilingue anglais-français: Données et modèles préliminaires [Comparing articulatory strategies in an English/French bilingual speaker: Data and preliminary models]. *Proceedings of XXX$^{ièmes}$ Journées d'Etude sur la Parole*, 448–456. https://hal.archives-ouvertes.fr/hal-01228883/document.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*(6), 493–503. doi: 10.1016/j.specom.2010.03.002

Bakst, S. (2016). Differences in the relationship between palate shape, articulation, and acoustics of American English /r/ and /s/. *UC Berkeley Phonology Lab Annual Report*, *12*, 216–224. https://escholarship.org/uc/item/0sj6b0zp.

Barras, W. (2008, April 23-25). *"We would say 'a bit Gretnarish' and we'd put an r in": Rhoticity and r-sandhi in East Lancashire speech* [Conference abstract]. The LEL Postgraduate Conference, Edinburgh, UK. https://pgc.lel.ed.ac.uk/archive/2008/abstracts/Will%20Barras.pdf.

Barras, W. (2010). *The sociophonology of rhoticity and r-sandhi in East Lancashire English* (PhD Thesis). University of Edinburgh, UK.

Barton, K. (2018). *MuMIn: Multi-Model Inference.* https://cran.r-project.org/web/packages/

MuMIn/index.html.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beal, J. C. (2009). "You're not from New York City, you're from Rotherham": Dialect and identity in British indie music. *Journal of English Linguistics*, *37*(3), 223–240. doi: 10.1177/0075424209340014

Bear, H. L., Harvey, R. W., Theobald, B.-J., & Lan, Y. (2014). Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In G. Bebis et al. (Eds.), *Advances in Visual Computing. ISVC 2014. Lecture Notes in Computer Science* (Vol. 8888, pp. 230–239). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-14364-4_22

Beckford Wassink, A., Wright, R. A., & Franklin, A. D. (2007). Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *Journal of Phonetics*, *35*(3), 363–379. doi: 10.1016/j.wocn.2006.07.002

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627. doi: 10.1006/jpho.2002.0177

Beddor, P. S., & Krakow, R. A. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America*, *106*(5), 2868–2887. doi: 10.1121/1.428111

Binnie, C. A., Jackson, P. L., & Montgomery. (1976). Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, *41*(4), 530–539. doi: 10.1044/jshd.4104.530

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer.* Version 6.0.50, retrieved from http://www.praat.org/.

Bohn, O.-S., & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, *40*(1), 109–128. doi: 10.1016/j.wocn.2011.08.002

Boyce, S. E., & Espy-Wilson, C. Y. (1997). Coarticulatory stability in American English /ɹ/. *The Journal of the Acoustical Society of America*, *101*(6), 3741–3753. doi: 10.1121/1.418333

Boyce, S. E., Hamilton, S. M., & Rivera-Campos, A. (2016). Acquiring rhoticity across languages: An ultrasound study of differentiating tongue movements. *Clinical Linguistics & Phonetics*, *30*(3-5), 174–201. doi: 10.3109/02699206.2015.1127999

Boyce, S. E., Tiede, M. K., Espy-Wilson, C. Y., & Groves-Wright, K. (2015). Diversity of tongue shapes for the American English rhotic liquid. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18$^{th}$ International Congress of Phonetic Sciences.* Glasgow, UK: The University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0847.pdf.

Bradlow, A. R. (2002). Confluent talker-and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology* (Vol. 7, pp. 241–273). New York, USA: Mouton de Gruyter.

Brooke, N. (1998). Computational aspects of visual speech: Machines that can speechread and simulate talking faces. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 109–1022). Hove, UK: Psychology Press.

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In F. Bell-Berti & L. Raphael (Eds.), *Producing speech: Contemporary issues. For Katherine Safford Harris* (pp. 19–33). New York, USA: AIP Press.

Brown, G. (1981). Consonant rounding in British English: The status of phonetic descriptions as historical data. In R. Asher & E. J. Henderson (Eds.), *Towards a history of phonetics* (pp. 67–76). Edinburgh University Press.

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M. K., & Perkell, J. S. (2011). The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research*, *54*(3), 727–739. doi: 10.1044/1092-4388(2010/09-0256)

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. doi: 10.1002/dev.20032

Burnham, D., Joeffry, S., & Rice, L. (2010a). Computer- and human-directed speech before and

after correction. In M. Tabain, J. Fletcher, D. Grayden, J. Hajek, & A. Butcher (Eds.), *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (pp. 13–17). Australia: The Australasian Speech Science and Technology Association Inc. https://assta.org/proceedings/sst/SST-10/SST2010/PDF/AUTHOR/ST100077.PDF.

Burnham, D., Joeffry, S., & Rice, L. (2010b). "d-o-e-s-not-c-o-m-p-u-t-e": Vowel hyperarticulation in speech to an auditory-visual avatar. *Proceedings of the 9th International Conference on Auditory-Visual Speech Processing (AVSP-2010)*. https://www.isca-speech.org/archive/avsp10/papers/av10_P18.pdf.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435–1435. doi: 10.1126/science.1069587

Burnham, D., & Lau, S. (1998). The effect of tonal information on auditory reliance in the McGurk effect. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98)* (p. 37-42). https://www.isca-speech.org/archive_open/archive_papers/avsp98/av98_037.pdf.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyperarticulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86. doi: 10.1016/j.jml.2015.12.009

Campbell, F., Gick, B., Wilson, I., & Vatikiotis-Bateson, E. (2010). Spatial and temporal properties of gestures in North American English /r/. *Language and Speech*, *53*(1), 49–69. doi: 10.1177/0023830909351209

Carr, P., & Durand, J. (2004). General American and New York City English. *La Tribune Internationale des Langues Vivantes*, *36*, 56–69.

Castellanos, A., Benedí, J.-M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, *20*(1-2), 23–35. doi: 10.1016/S0167-6393(96)00042-8

Catford, J. C. (1977). *Fundamental problems in phonetics*. Edinburgh University Press.

Catford, J. C. (1988). *A practical introduction to phonetics*. Oxford, UK: Clarendon Press.

Catford, J. C. (2001). On Rs, rhotacism and paleophony. *Journal of the International Phonetic Association*, *31*(2), 171–185. doi: 10.1017/S0025100301002018

Cattelain, T., Garnier, M., Savariaux, C., Gerber, S., & Perrier, P. (2018). Analyse électromyographique de la production des plosives labiales : Enjeux méthodologiques [Electromyographic analysis of the production of labial plosives: Methodological issues]. *Proceedings of XXXII$^{ièmes}$ Journées d'Etude sur la Parole*, 107–115. https://www.isca-speech.org/archive/JEP_2018/pdfs/192698.pdf.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science* (Vol. 11211, pp. 833–851). Cham, Switzerland: Springer. doi: 10.1007/978-3-030-01234-2_49

Chen, S., Tiede, M. K., & Whalen, D. (2017). Dynamic transitional differences between American English bunched and retroflex /ɹ/: Articulatory and acoustic measures. *Proceedings of the 11$^{th}$ International Seminar on Speech Production.*

Cheyne, H. A., Kalgaonkar, K., Clements, M., & Zurek, P. (2009). Talker-to-listener distance effects on speech production and perception. *The Journal of the Acoustical Society of America*, *126*(4), 2052–2060. doi: 10.1121/1.3205400

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure.* Kaiseikan, Tokyo.

Chitoran, I. (2012). The nature of historical change. In A. C. Cohn, C. Fougeron, & M. K. Huffman (Eds.), *Handbook of laboratory phonology* (pp. 311–321). Oxford University Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York, USA: Harper and Row.

Cialdella, L., Kabakoff, H., Preston, J. L., Dugan, S., Spencer, C., Boyce, S., … McAllister, T. (2020). Auditory-perceptual acuity in rhotic misarticulation: Baseline characteristics and treatment response. *Clinical Linguistics & Phonetics*, *0*(0), 1–24. doi: 10.1080/02699206.2020.1739749

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409–412. doi: 10.3758/BF03203962

Colin, C., Radeau, M., Deltenre, P., Demolin, D., & Soquet, A. (2002). The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *European Journal of*

*Cognitive Psychology*, *14*(4), 475–491. doi: 10.1080/09541440143000203

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, *28*(2), 543–571. doi: 10.1016/j.csl.2013.08.003

Corballis, M. C. (2003). From mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences*, *26*(2), 199–208. doi: 10.1017/S0140525X03000062

Corballis, M. C. (2014). *The recursive mind: The origins of human language, thought, and civilization* (updated ed.). Princeton, NJ, USA; Oxford, UK: Princeton University Press.

Costa, M. G., Campos, J. P., de Aquino e Aquino, G., de Albuquerque Pereira, W. C., & Costa Filho, C. F. F. (2019). Evaluating the performance of convolutional neural networks with direct acyclic graph architectures in automatic segmentation of breast lesion in US images. *BMC Medical Imaging*, *19*(1), 85. doi: 10.1186/s12880-019-0389-2

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Cvejic, E., Kim, J., Davis, C., & Gibert, G. (2010). Prosody for the eyes: Quantifying visual prosody using guided principal component analysis. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11$^{th}$ Annual Conference of the International Speech Communication Association (Interspeech 2010)* (pp. 1433–1436). https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1433.pdf.

Dalcher, C. V., Knight, R.-A., & Jones, M. J. (2008). Cue switching in the perception of approximants: Evidence from two English dialects. *The Journal of the Acoustical Society of America*, *123*(5), 3319. doi: 10.1121/1.2933789

Dalston, R. M. (1975). Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults. *The Journal of the Acoustical Society of America*, *57*(2), 462–469. doi: 10.1121/1.380469

Dancer, J., Krain, M., Thompson, C., Davis, P., & et al. (1994). A cross-sectional investigation of speechreading in adults: Effects of age, gender, practice, and education. *The Volta Review*, *96*(1), 31–40.

Dediu, D., & Moisik, S. R. (2019). Pushes and pulls from below: Anatomical variation, articulation and sound change. *Glossa: A Journal of General Linguistics*, *4*(1), 7. doi: 10.5334/gjgl.646

Delattre, C., Pierre, & Freeman, D. C. (1968). A dialect study of American r's by X-ray motion picture. *Linguistics*, *6*(44), 29–68. doi: 10.1515/ling.1968.6.44.29

Denes, P. B., & Pinson, E. N. (1993). *The speech chain: The physics and biology of spoken language.* Long Grove, IL, USA: Waveland Press, Inc.

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology*, *66*(1), 85–110. doi: 10.1006/jecp.1997.2379

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187–203. doi: 10.1002/dev.20033

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144. doi: 10.1207/s15326969eco0102_2

Docherty, G., & Foulkes, P. (2001). Variability in (r) production – Instrumental perspectives. In H. Van de Velde & R. van Hout (Eds.), *'R-Atics: Sociolinguistic, phonetic and phonological characteristics of* /r/ (pp. 173–184). Brussels, Belgium: Université Libre de Bruxelles.

Dodd, B., Holm, A., Hua, Z., & Crosbie, S. (2003). Phonological development: A normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, *17*(8), 617–643. doi: 10.1080/0269920031000111348

Dohen, M. (2009). Speech through the ear, the eye, the mouth and the hand. In A. Esposito, A. Hussain, M. Marinaro, & R. Martone (Eds.), *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 24–39). Berlin/Heidelberg, Germany: Springer.

Dyson, Alice Tanner. (1988). Phonetic inventories of 2- and 3-year-old children. *Journal of Speech and Hearing Disorders*, *53*(1), 89–93. doi: 10.1044/jshd.5301.89

Ehrlich, S., & Avery, P. (2013). *Teaching American English pronunciation – Oxford handbooks for language teachers.* Oxford University Press.

Epstein, M. A., & Stone, M. (2005). The tongue stops here: Ultrasound imaging of the palate. *The Journal of the Acoustical Society of America*, *118*(4), 2128–2131. doi: 10.1121/1.2031977

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing*

*Disorders*, *40*(4), 481–492. doi: 10.1044/jshd.4004.481

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(SI), 253–260. doi: 10.1016/S0095-4470(11)00055-6

Eskes, M., van Alphen, M. J., Balm, A. J., Smeele, L. E., Brandsma, D., & van der Heijden, F. (2017). Predicting 3D lip shapes using facial surface EMG. *PLOS ONE*, *12*(4), e0175025. doi: 10.1371/journal.pone.0175025

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. *The Journal of the Acoustical Society of America*, *92*(2), 736–757. doi: 10.1121/1.403998

Espy-Wilson, C. Y., & Boyce, S. E. (1999). A simple tube model for American English /r/. In J. J. Ohala, Y. Hasegawa, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14$^{th}$ International Congress of Phonetic Sciences* (pp. 2137–2140). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_2137.pdf.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America*, *108*(1), 343–356. doi: 10.1121/1.429469

Espy-Wilson, C. Y., & Boyce, S. E. (1994). Acoustic differences between "bunched" and "retroflex" variants of American English /r/. *The Journal of the Acoustical Society of America*, *95*(5), 2823–2823. doi: 10.1121/1.409691

Fabricius, A. (2007). Vowel formants and angle measurements in diachronic sociophonetic studies: FOOT-fronting in RP. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16$^{th}$ International Congress of Phonetic Sciences* (pp. 1477–1480). http://www.icphs2007.de/conference/Papers/1087/1087.pdf.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Netherlands: Mouton.

Fant, G. (1989). Quantal theory and features. *Journal of Phonetics*, *17*(1), 79–86. doi: 10.1016/S0095-4470(19)31525-6

Ferragne, E. (2019, June 4). *Phonetics and artificial intelligence: Ready for the paradigm shift?* [Conference abstract]. The Phonology of Contemporary English Conference, Aix-en-

Provence, France.

Ferragne, E., Gendrot, C., & Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19$^{th}$ International Congress of Phonetic Sciences* (pp. 790–794). Canberra, Australia: Australasian Speech Science and Technology Association Inc. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_839.pdf.

Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, *40*(1), 1–34. doi: 10.1017/S0025100309990247

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. doi: 10.1044/jshr.1104.796

Fitzpatrick, M., Kim, J., & Davis, C. (2015). The effect of seeing the interlocutor on auditory and visual speech production in noise. *Speech Communication*, *74*, 37–51. doi: 10.1016/j.specom.2015.08.001

Fitzroy, A. B., Ugolini, M., Munoz, M., Zobel, B. H., Sherwood, M., & Sanders, L. D. (2018). Attention modulates early auditory processing at a real cocktail party. *Language, Cognition and Neuroscience*, *0*(0), 1–17. doi: 10.1080/23273798.2018.1492002

Fixmer, E., & Hawkins, S. (1998). The influence of quality of information on the McGurk effect. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing* (pp. 27–32). https://www.isca-speech.org/archive_open/archive_papers/avsp98/av98_027.pdf.

Flemming, E. S. (2013). *Auditory representations in phonology*. New York, USA: Routledge. doi: 10.4324/9781315054803

Folkins, J. (1978). Lower lip displacement during in-vivo stimulation of human labial muscles. *Archives of Oral Biology*, *23*(3), 195–202. doi: 10.1016/0003-9969(78)90216-9

Fong, R. C., Scheirer, W. J., & Cox, D. D. (2018). Using human brain activity to guide machine learning. *Scientific Reports*, *8*(1), 1–10. doi: 10.1038/s41598-018-23618-6

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains.

*The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. doi: 10.1121/1.418332

Foulkes, P. (1997). English [r]-sandhi – A sociolinguistic perspective. *Histoire Épistémologie Langage*, *19*(1), 73–96. doi: 10.3406/hel.1997.2573

Foulkes, P., & Docherty, G. J. (2000). Another chapter in the story of /r/: 'Labiodental' variants in British English. *Journal of Sociolinguistics*, *4*(1), 30–59. doi: 10.1111/1467-9481.00102

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, *14*(1), 3–28. doi: 10.1016/S0095-4470(19)30607-2

Franks, R. J., & Kimble, J. (1972). The confusion of English consonant clusters in lipreading. *Journal of Speech and Hearing Research*, *15*(3), 474–482. doi: 10.1044/jshr.1503.474

Gagné, J.-P., Rochette, A.-J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication*, *37*(3), 213–230. doi: 10.1016/S0167-6393(01)00012-7

Garnier, M., Bailly, L., Dohen, M., Welby, P., & Loevenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: Global effects on the utterance. *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, 2246–2249. https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1862.pdf.

Garnier, M., Heinrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *Journal of Speech, Language, and Hearing Research*, *53*(3), 588–608. doi: 10.1044/1092-4388(2009/08-0138)

Garnier, M., Ménard, L., & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America*, *144*(2), 1059–1074. doi: 10.1121/1.5051321

Garnier, M., Ménard, L., & Richard, G. (2012). Effect of being seen on the production of visible speech cues. a pilot study on Lombard speech. *Proceedings of 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 611–614. https://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0611.pdf.

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics*, *44*, 83–95. doi: 10.1016/j.wocn.2014.02.006

Gick, B. (1999). A gesture-based account of intrusive consonants in English. *Phonology*, *16*(1),

29–54. doi: 10.1017/S0952675799003693

Gick, B. (2002a). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121. doi: 10.1017/S0025100302001007

Gick, B. (2002b). An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica*, *59*(1), 38–48. doi: 10.1159/000056204

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. In J. Hansen & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 309–322). Amsterdam, Netherlands: John Benjamins.

Gick, B., Chiu, C., Widing, E., Roewer-Despres, F., Mayer, C., Fels, S., & Stavness, I. (2019). Quantal biomechanical effects in speech postures of the lips. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19$^{th}$ International Congress of Phonetic Sciences* (pp. 1749–1753). Canberra, Australia: Australasian Speech Science and Technology Association Inc. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1798.pdf.

Gick, B., Kang, M. A., & Whalen, D. H. (2000). MRI and X-ray evidence for commonality in the dorsal articulations of English vowels and liquids. *Proceedings of the 5$^{th}$ Seminar on Speech Production: Models and data.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.514.4315&rep=rep1&type=pdf.

Giegerich, H. (1999). *Lexical Strata in English: Morphological causes, phonological effects.* Cambridge University Press.

Giles, H., & Smith, P. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. N. St. Clair (Eds.), *Language and social psychology* (pp. 45–65). Oxford, UK: Blackwell.

Giles, S., & Moll, K. (1975). Cinefluorographic study of selected allophones of English /l/. *Phonetica*, *31*, 206–227.

Gimson, A. (1980). *An introduction to the pronunciation of English.* London, UK: Arnold.

Graf, P. H., Costatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the 5$^{th}$ IEEE International Conference of Automatic Face Gesture Recognition.* doi: 10.1109/AFGR.2002.1004186

Granström, B., House, D., & Lundeberg, M. (1999). Prosodic cues in multimodal speech perception. In J. J. Ohala, Y. Hasegawa, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14$^{th}$ International Congress of Phonetic Sciences* (pp. 655–658). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0655.pdf.

Grant, K. W., & Seitz, P.-F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, *104*(4), 2438–2450. doi: 10.1121/1.423751

Grant, K. W., Walden, B. E., & Seitz, P.-F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677–2690. doi: 10.1121/1.422788

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York, USA: Wiley.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America*, *105*(5), 2854–2865. doi: 10.1121/1.426900

Hagiwara, R. (1995). Acoustic realizations of American /r/ as produced by women and men. *UCLA Working Papers in Phonetics*, *90*, 1–187. http://escholarship.org/uc/item/8779b7gq.

Hallé, P. A., Best, C. T., & Levitt, A. (1999). Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics*, *27*(3), 281–306. doi: 10.1006/jpho.1999.0097

Hamann, S. (2002). Retroflexion and retraction revised. *ZAS Working Papers in Linguistics*, *28*, 13–25. https://core.ac.uk/download/pdf/19210267.pdf.

Hamann, S. (2003). *The phonetics and phonology of retroflexes* (PhD Thesis). LOT, Utrecht, Netherlands.

Hancock, M. (2003). *English pronunciation in use.* Cambridge University Press.

Harrington, J. (2010). Acoustic phonetics. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.),

*The handbook of phonetic sciences* (Second ed., pp. 81–129). Oxford, UK: Blackwell.

Harrington, J., & Cassidy, S. (1999). *Techniques in speech acoustics*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, *123*(5), 2825–2835. doi: 10.1121/1.2897042

Harrington, J., Kleber, F., & Reubold, U. (2011). The contributions of the lips and the tongue to the diachronic fronting of high back vowels in Standard Southern British English. *Journal of the International Phonetic Association*, *41*(2), 137–156. doi: 10.1017/S0025100310000265

Havenhill, J. (2018). *Constraints on articulatory variability: Audiovisual perception of lip rounding* (PhD Thesis). Georgetown University, Washington, DC.

Havenhill, J., & Do, Y. (2018). Visual speech perception cues constrain patterns of articulatory variation and sound change. *Frontiers in Psychology*, *9*(728). doi: 10.3389/fpsyg.2018.00728

Hay, J., & Sudbury, A. (2005). How rhoticity became /r/-sandhi. *Language*, *81*(4), 799–823. doi: 10.1353/lan.2005.0175

Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., & Diehl, R. L. (2006). Enhanced contrast for vowels in utterance focus: A cross-language study. *The Journal of the Acoustical Society of America*, *119*(5), 3022–3033. doi: 10.1121/1.2184226

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740–1751. doi: 10.1121/1.2166611

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

Hecht, D., & Reiner, M. (2009). Sensory dominance in combinations of audio, visual and haptic stimuli. *Experimental Brain Research*, *193*(2), 307–314. doi: 10.1007/s00221-008-1626-z

Heffner, R.-M. (1950). *General phonetics*. Madison, WI, USA: University of Wisconsin Press.

Heselwood, B., & Plug, L. (2011). The role of F2 and F3 in the perception of rhoticity: Evidence from listening experiments. *Proceedings of the 17$^{th}$ International Congress of Phonetic*

*Sciences*, 867–870. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Heselwood/Heselwood.pdf.

Heyne, M., Wang, X., Derrick, D., Dorreen, K., & Watson, K. (2018). The articulation of /ɹ/ in New Zealand English. *Journal of the International Phonetic Association*, 1–23. doi: 10.1017/S0025100318000324

Honda, K., Kurita, T., Kakita, Y., & Maeda, S. (1995). Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*, *23*(1), 243–254. doi: 10.1016/S0095-4470(95)80046-8

Howell, P., & Bonnett, C. (1997). Speaking clearly for the hearing impaired: Intelligibility differences between clear and less clear speakers. *International Journal of Language & Communication Disorders*, *32*(1), 89–97. doi: 10.3109/13682829709021462

Irwin, J. R., Frost, S. J., Mencl, W. E., Chen, H., & Fowler, C. A. (2011). Functional activation for imitation of seen and heard speech. *Journal of Neurolinguistics*, *24*(6), 611–618. doi: 10.1016/j.jneuroling.2011.05.001

Johnson, K. (2012). *Acoustic and auditory phonetics* (3rd ed.). Wiley-Blackwell.

Jones, D. (1956). *The pronunciation of English.* Cambridge University Press.

Jones, D. (1972). *An outline of English phonetics* (9th ed.). Cambridge University Press.

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1367–1377. doi: 10.1044/1092-4388(2003/106)

Jordan, T. R., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech*, *43*(1), 107–124. doi: 10.1177/00238309000430010401

Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524. doi: 10.1121/1.405631

Kenyon, J. (1940). *American pronunciation.* Ann Arbor, MI, USA: George Wahr.

Kerswill, P. (1996). Children, adolescents, and language change. *Language Variation and Change*, *8*(2), 177–202. doi: 10.1017/S0954394500001137

Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception*, *40*, 853–862. doi: 10.1068/p6941

King, H., & Ferragne, E. (2018, June 27-29). /u/*-fronting in English: How phonetically accurate should phonological labels be?* [Conference abstract]. 16èmes Rencontres Du Réseau Français de Phonologie, Paris, France. https://www.sfl.cnrs.fr/sites/default/files/images/abstracts-general-session_0.pdf#page=21.

King, H., & Ferragne, E. (2019). The contribution of lip protrusion to Anglo-English /r/: Evidence from hyper-and non-hyperarticulated speech. *Proceedings of Interspeech 2019*, 3322–3326. doi: 10.21437/Interspeech.2019-2851

King, H., & Ferragne, E. (2020). Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English. *Journal of Phonetics*, *80*, 100978. doi: 10.1016/j.wocn.2020.100978

Klause, F., Stone, S., & Birkholz, P. (2017). A head-mounted camera system for the measurement of lip protrusion and opening during speech production. In J. Trouvain, I. Steiner, & B. Möbius (Eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017* (pp. 145–151). Dresden, Germany: TUD press, Dresden.

Klein, H. B., Grigos, M. I., Byun, T. M., & Davidson, L. (2012). The relationship between inexperienced listeners' perceptions and acoustic correlates of children's /r/ productions. *Clinical Linguistics & Phonetics*, *26*(7), 628–645. doi: 10.3109/02699206.2012.682695

Knight, R.-A., Dalcher, C. V., & Jones, M. J. (2007). A real-time case study of rhotic acquisition in Southern British English. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16$^{th}$ International Congress of Phonetic Sciences* (pp. 1581–1584). http://www.icphs2007.de/conference/Papers/1352/1352.pdf.

Kochetov, A. (2020). Research methods in articulatory phonetics I: Introduction and studying oral gestures. *Language and Linguistics Compass*, e12368. doi: 10.1111/lnc3.12368

Koppen, C., & Spence, C. (2007). Seeing the light: Exploring the Colavita visual dominance effect. *Experimental Brain Research*, *180*(4), 737–754. doi: 10.1007/s00221-007-0894-3

Krämer, M. (2012). *Underlying representations.* Cambridge, UK; New York, USA: Cambridge University Press.

Krause, J. C., & Braida, L. D. (2003). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, *115*(1), 362–378.

doi: 10.1121/1.1635842

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review*, *84*(4), 219–225.

Kuehn, D. P., & Tomblin, J. B. (1977). A cineradiography investigation of children's w/r substitutions. *Journal of Speech and Hearing Disorders*, *42*(4), 462-473. doi: 10.1044/jshd.4204.462

Kuhl, P., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141. doi: 10.1126/science.7146899

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., … Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. doi: 10.1126/science.277.5326.684

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). doi: 10.18637/jss.v082.i13

van Laarhoven, T., Keetels, M., Schakel, L., & Vroomen, J. (2018). Audio-visual speech in noise perception in dyslexia. *Developmental Science*, *21*(1). doi: 10.1111/desc.12504

Labov, W., Ash, S., & Boberg, C. (2008). *Atlas of North American English: Phonetics, phonology and sound change*. Berlin, Germany: Mouton de Gruyter.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago, IL, USA: The University of Chicago Press.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants* (3rd ed.). Chichester, UK: Wiley-Blackwell.

Ladefoged, P., Epstein, M., & Hacopian, N. (2002). Dissection manual for students of speech. *UCLA Working Papers in Phonetics*, *102*. https://linguistics.ucla.edu/people/ladefoge/manual.htm.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, UK: Blackwell.

Lallouache, M. T. (1991). *Un poste « visage-parole » couleur. Acquisition et traitement automatique des contours des lèvres* [A "face-speech" interface. Automatic acquisition and processing of labial contours] (PhD Thesis). ENSERG, Grenoble, France.

Lalonde, K., & Frush Holt, R. (2015). Preschoolers benefit from visually salient speech cues. *Jour-*

*nal of Speech, Language, and Hearing Research*, *58*(1), 135–150. doi: 10.1044/2014_JSLHR-H-13-0343

Lalonde, K., & Werner, L. A. (2019). Infants and adults use visual cues to improve detection and discrimination of speech in noise. *Journal of Speech, Language, and Hearing Research : JSLHR*, *62*(10), 3860–3875. doi: 10.1044/2019_JSLHR-H-19-0106

Laver, J. (1980). *The phonetic description of voice quality: Cambridge studies in linguistics*. Cambridge, UK: Cambridge University Press.

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2011). The social stratification of tongue shape for postvocalic /r/ in Scottish English. *Journal of Sociolinguistics*, *15*(2), 256–268. doi: 10.1111/j.1467-9841.2011.00464.x

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2013). Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation. *Journal of Phonetics*, *41*(3), 198–210. doi: 10.1016/j.wocn.2013.01.004

Lawson, E., Stuart-Smith, J., & Rodger, L. (2019). A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes. *The Journal of the Acoustical Society of America*, *146*(6), 4363–4381. doi: 10.1121/1.5139215

Lawson, E., Stuart-Smith, J., & Scobbie, J. M. (2014). A mimicry study of adaptation towards socially-salient tongue shape variants. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 12. https://repository.upenn.edu/pwpl/vol20/iss2/12.

Lawson, E., Stuart-Smith, J., & Scobbie, J. M. (2018). The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *The Journal of the Acoustical Society of America*, *143*(3), 1646–1657. doi: 10.1121/1.5027833

Lawson, E., Stuart-Smith, J., Scobbie, J. M., & Nakai, S. (2018). *Dynamic Dialects: An Articulatory Web Resource for the Study of Accents*. University of Glasgow. https://www.dynamicdialects.ac.uk/.

Lawson, E., Stuart-Smith, J., Scobbie, J. M., Yaeger-Dror, M., & Maclagan, M. (2010). Analyzing liquids. In M. De Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A student's guide* (pp. 72–86). London, UK: Routledge.

Lee, A. (2000). *Virtual Dub*. Version 1.10.4. http://www.virtualdub.org.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455–1468. doi: 10.1121/1.426686

Leemann, A., Kolly, M.-J., & Britain, D. (2018). The English dialects app: The creation of a crowdsourced dialect corpus. *Ampersand*, *5*, 1–17. doi: 10.1016/j.amper.2017.11.001

Lehiste, I. (1962). *Acoustical characteristics of selected English consonants.* Ann Arbor, MI, USA: University of Michigan Communication Sciences Laboratory.

Lennon, R., Smith, R., & Stuart-Smith, J. (2015). An acoustic investigation of postvocalic /r/ variants in two sociolects of Glaswegian. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK: The University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1019.pdf.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. doi: 10.18637/jss.v069.i01

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. doi: 10.1016/0010-0277(85)90021-6

Lilly, R., & Viel, M. (1977). *La prononciation de l'anglais: Règles phonologiques et exercices de transcription* [The pronunciation of English: Phonologicial rules and transcription exercises]. Paris, France: Hachette.

Lin, S., & Demuth, K. (2013). The gradual acquisition of English /l/. In S. Baiz, N. Goldman, & R. Hawkes (Eds.), *Proceedings of the 37th annual Boston University Conference on Language Development* (pp. 206–218). Somerville, MA, USA: Cascadilla Press.

Lin, S., & Demuth, K. (2015). Children's acquisition of English onset and coda /l/: Articulatory evidence. *Journal of Speech, Language, and Hearing Research*, *58*(1), 13–27. doi: 10.1044/2014_JSLHR-S-14-0041

Lindau, M. (1978). Vowel features. *Language*, *54*(3), 541–563. doi: 10.1353/lan.1978.0066

Lindau, M. (1985). The story of /r/. In V. Fromkin (Ed.), *Phonetic linguistics: Essay in honor of Peter Ladefoged.* Orlando, FL, USA: Academic Press.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcas-

tle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht, The Netherlands: Springer. doi: 10.1007/978-94-009-2037-8_16

Lindblom, B., Brownlee, S., Davis, B., & Moon, S. J. (1992). Speech transforms. *Speech Communication*, *11*(4), 357–368. doi: 10.1016/0167-6393(92)90041-5

Lindblom, B., Sundberg, J., Branderud, P., Djamshidpey, H., & Granqvist, S. (2010). The Gunnar Fant legacy in the study of vocal acoustics. *Proceedings of 10ème Congrès Français d'Acoustique*. https://hal.archives-ouvertes.fr/hal-00539775/document.

Lindley, N., & Lawson, E. (2016, March-April 30-1). *An articulatory investigation of Anglo-English prevocalic* /r/ [Conference abstract]. BAAP Colloquim, Lancaster, UK. http://wp.lancs.ac.uk/phonetics/files/2015/08/BAAP_abstracts.pdf#page=49.

Linker, W. (1982). Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study. *UCLA Working Papers in Phonetics*, *56*. https://escholarship.org/uc/item/0wq546xq.

Lisker, L. (1957). Minimal cues for separating /w, r, l, y/ in intervocalic position. *Word*, *13*(2), 256–267. doi: 10.1080/00437956.1957.11659637

Llamas, C. (1998). Language variation and innovation in Middlesborough: A pilot study. *Leeds Working Papers in Linguistics and Phonetics*, *6*, 97–114.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608. doi: 10.1121/1.1912396

Lombard, E. (1911). Le signe de l'élévation de la voix [The sign of the elevation of the voice]. *Annales des Maladies de l'Oreille et du Larynx*, *37*(2), 101–119.

Lüdecke, D. (2018). *sjPlot – Data visualization for statistics in social science.* doi: 10.5281/zenodo.1310947

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (Second ed.). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Maddieson, I. (1984). *Pattern of sounds.* Cambridge, UK: Cambridge University Press.

Maddieson, I. (2008). Chapter 11: Front rounded vowels. In M. Dryer S, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online.* Munich, Germany: Max Plank Digital Library. http://wals.info/feature/description/11.

Maddieson, I., & Precoda, K. (1989). Updating UPSID. *The Journal of the Acoustical Society of America*, *86*(S1), S19-S19. doi: 10.1121/1.2027403

Magloughlin, L. (2016). Accounting for variability in North American English /r/: Evidence from children's articulation. *Journal of Phonetics*, *54*, 51–67. doi: 10.1016/j.wocn.2015.07.007

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*(9), 2581–2586. doi: 10.1007/s00221-015-4324-7

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962-3973. doi: 10.1121/1.2990715

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228. doi: 10.3758/BF03204377

Marchal, A. (2009). *From speech physiology to linguistic phonetics* (Vol. 145). London, UK: Wiley.

Marieb, E. N., & Hoehn, K. (2007). *Human anatomy & physiology* (7th ed.). San Francisco, CA, USA: Pearson education.

Marks, J. (2007). *English pronunciation in use (Elementary).* Cambridge University Press.

Marsden, S. (2006). A sociophonetic study of labiodental /r/ in Leeds. *Leeds Working Papers in Linguistics and Phonetics*(11), 153–172. https://www.latl.leeds.ac.uk/wp-content/uploads/sites/49/2019/05/Marsden_2006.pdf.

Masapollo, M., Polka, L., & Ménard, L. (2017). A universal bias in adult vowel perception – By ear or by eye. *Cognition*, *166*, 358–370. doi: 10.1016/j.cognition.2017.06.001

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*(5), 1777–1788. doi: 10.2307/1129925

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NK, USA: Lawrence Erlbaum Associates.

Mathworks. (2019). *MATLAB deep learning toolbox R2019a.* Natick, MA, USA.

Mathworks. (2020a). *MATLAB computer vision toolbox R2020a.* Natick, MA, USA.

Mathworks. (2020b). *MATLAB image processing toolbox R2020a.* Natick, MA, USA.

Mattheyses, W., & Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, *66*, 182–217. doi: 10.1016/j.specom.2014.11.001

Matthies, M. L., Guenther, F. H., Denny, M., Perkell, J. S., Burton, E., Vick, J., … Zandipour, M. (2008). Perception and production of /r/ allophones improve with hearing from a cochlear implant. *The Journal of the Acoustical Society of America*, *124*(5), 3191–3202. doi: 10.1121/1.2987427

Mayr, R. (2010). What exactly is a front rounded vowel? An acoustic and articulatory investigation of the NURSE vowel in South Wales English. *Journal of the International Phonetic Association*, *40*(1), 93–112. doi: 10.1017/S0025100309990272

McCloy, D. (2013). *Mix speech with noise.* Praat script licensed under the GNU General Public Licence v3.0, retrieved from https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat.

McGowan, R. S., Nittrouer, S., & Manning, C. J. (2004). Development of [ɪ] in young, Midwestern, American children. *The Journal of the Acoustical Society of America*, *115*(2), 871–884. doi: 10.1121/1.1642624

McGuire, G., & Babel, M. (2012). A cross-modal account for synchronic and diachronic patterns of /f/ and /θ/ in English. *Laboratory Phonology*, *3*(2), 251–272. doi: 10.1515/lp-2012-0014

McGurk, H. (1981). *The Auditory-Visual Perception of Speech* [Unpublished Manuscript]. University of Surrey, UK.

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi: 10.1038/264746a0

McMahon, A., Foulkes, P., & Tollfree, L. (1994). Gestural representation and lexical phonology. *Phonology*, *11*(2), 277–316. doi: 10.1017/S0952675700001974

Ménard, L., Dupont, S., Baum, S. R., & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America*, *126*(3), 1406–1414. doi: 10.1121/1.3158930

Ménard, L., Schwartz, J.-L., Boë, L.-J., & Aubin, J. (2007). Articulatory-acoustic relationships during vocal tract growth for French vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, *35*(1), 1–19. doi: 10.1016/j.wocn.2006.01.003

Ménard, L., Trudeau-Fisette, P., Côté, D., & Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers. *PLOS ONE*, *11*(9), e0160088. doi: 10.1371/journal.pone.0160088

Meyer, J. (2005). Whistled speech: A natural phonetic description of languages adapted to human perception and to the acoustical environment. *Proceedings of Interspeech 2005*, 49–52. https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_0049.pdf.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language*, *92*(1), 101–140. doi: 10.1353/lan.2016.0019

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. doi: 10.1121/1.1907526

Milroy, L., & Gordon, M. (2008). The concept of social network. In *Sociolinguistics: Method and interpretation* (pp. 116–133). Oxford, UK: John Wiley & Sons.

Munhall, K., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. doi: 10.1007/s00426-005-0031-5

Nogita, A., & Yamane, N. (2018). Tokyo Japanese /u/ to be unrounded or rounded?: Redefining roundness, compression, and protrusion. *Proceedings from the Phonology Forum*. doi: 10.13140/RG.2.2.25342.84800

Nogita, A., Yamane, N., & Bird, S. (2013, November 6-8). *The Japanese Unrounded Back Vowel* [ɯ] *Is in Fact Rounded Central/Front* [ʉ-ʏ] [Conference abstract]. Ultrafest VI, Edinburgh, UK. https://www.qmu.ac.uk/media/5703/ultrafest-abstract-booklet-2013.pdf#page=43.

Noiray, A., Cathiard, M.-A., Ménard, L., & Abry, C. (2011). Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *The Journal of the Acoustical Society of America*, *129*(1), 340–349. doi: 10.1121/1.3518452

Noiray, A., Ries, J., & Tiede, M. (2015, December 8-10). *Sonographic & Optical Linguo-Labial Articulation Recording System (SOLLAR)* [Conference abstract]. Ultrafest VII, Hong Kong. http://www.ultrafest2015.hku.hk/docs/ABSTRACT%20BOOK.pdf#page=19.

O'Connor, J. D. (1967). *Better English pronunciation.* Cambridge, UK: Cambridge University Press.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, *13*(1), 24–43. doi: 10.1080/00437956.1957.11659626

O'Dwyer, N. J., Quinn, P. T., Guitar, B. E., Andrews, G., & Neilson, P. D. (1981). Procedures for verification of electrode placement in EMG studies of orofacial and mandibular muscles. *Journal of Speech, Language, and Hearing Research*, *24*(2), 273–288. doi: 10.1044/jshr.2402.273

Ohala, J. J. (1981). The listener as a source of sound change. In C. Masek & R. Hendrick (Eds.), *Papers from the parasession on language and behavior* (pp. 178–203). Chicago, IL, USA: Chicago Linguistic Society.

Ohala, J. J. (1985). Around flat. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 223–241). Orlando, FL, USA: Academic.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*(3), 1718–1725. doi: 10.1121/1.414696

Ong, D., & Stone, M. (1998). Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics. *Phonoscope*, *1*(1), 1–13.

Orton, H., & Dieth, E. (1962). *Survey of English dialects* (Vol. 1). Leeds, UK: E.J. Arnold & Son.

Oviatt, S., Levow, G., MacEachern, M., & Kuhn, K. (1996). Modeling hyperarticulate speech during human-computer error resolution. *Proceeding of 4$^{th}$ International Conference on Spoken Language Processing (ICSLP'96)*, *2*, 801-804 vol.2. doi: 10.1109/ICSLP.1996.607722

Pasquereau, J. (2018). Phonological degrees of labiality. *Language*, *94*(4), e216-e265. doi: 10.1353/lan.2018.0066

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*(2), 237–247. doi:

10.1016/S0163-6383(99)00003-X

Peelle, J. E. (2019). The neural basis for auditory and audiovisual speech perception. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 193–216). New York, USA: Routledge.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 169–181. doi: 10.1016/j.cortex.2015.03.006

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Pelegrín-García, D., Smits, B., Brunskog, J., & Jeong, C.-H. (2011). Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical Society of America*, *129*(4), 1981–1990. doi: 10.1121/1.3552881

Pereira, Y. I. (2013). Perception of English vowels and use of visual cues by learners of English and English native speakers. *Proceedings of Meetings on Acoustics*, *19*(1), 060120. doi: 10.1121/1.4800682

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study. *The Journal of the Acoustical Society of America*, *93*(5), 2948–2961. doi: 10.1121/1.405814

Perrier, P., & Fuchs, S. (2015). Motor equivalence in speech production. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 225–247). Wiley Blackwell.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *28*(1), 96–103. doi: 10.1044/jshr.2801.96

Piercy, C. (2012). A transatlantic cross-dialectal comparison of non-prevocalic /r/. *University of Pennsylvania Working Papers in Linguistics*, *18*(2), Article 10. https://repository.upenn.edu/pwpl/vol18/iss2/10.

Proctor, M., Walker, R., Smith, C., Szalay, T., Goldstein, L., & Narayanan, S. (2019). Articulatory characterization of English liquid-final rimes. *Journal of Phonetics*, *77*, 100921. doi:

10.1016/j.wocn.2019.100921

Queen Margaret University. (2017). *Gender action plan.* https://www.qmu.ac.uk/media/6789/gender-action-plan_july-2017.pdf.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. https://www.R-project.org/. Vienna, Austria.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London, UK: Lawrence Erlbaum Associates.

Roach, P. (1983). *English phonetics and phonology: A practical course* (2nd ed.). Cambridge University Press.

Rosenblum, L. D. (2008a). Primacy of multimodal speech perception. In D. B. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (pp. 51–78). Malden, MA, USA: Blackwell Publishing.

Rosenblum, L. D. (2008b). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. doi: 10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, *52*(4), 461–473. doi: 10.3758/BF03206706

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *The European Journal of Neuroscience*, *33*(12), 2329–2337. doi: 10.1111/j.1460-9568.2011.07685.x

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. doi: 10.1093/cercor/bhl024

Rubin-Spitz, J., McGarr, N. S., & Youdelman, K. (1986). Perception of stress contrasts by the hearing impaired. *The Journal of the Acoustical Society of America*, *79*(S1), S10-S10. doi: 10.1121/1.2023065

Saitoh, T., & Konishi, R. (2010). Profile lip reading for vowel and word recognition. *Proceedings*

*of the 20th International Conference on Pattern Recognition*, 1356–1359.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*(1-2), 75–87. doi: 10.1016/S0167-6393(98)00051-X

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. doi: 10.2307/3002019

Savariaux, C., Perrier, P., & Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428–2442. doi: 10.1121/1.413277

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *The Journal of the Acoustical Society of America*, *121*(5), 3044–3044. doi: 10.1121/1.4781735

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, *41*(3-4), 249–263. doi: 10.1016/j.wocn.2013.03.007

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. doi: 10.1038/nmeth.2089

Schwartz, J.-L., Abry, C., Boë, L.-J., Ménard, L., & Vallée, N. (2005). Asymmetries in vowel perception, in the context of the dispersion-focalisation theory. *Speech Communication*, *45*(4), 425–434. doi: 10.1016/j.specom.2004.12.001

Scobbie, J. M. (2006). (R) as a variable. In K. Brown (Ed.), *The encyclopaedia of language and linguistics* (Second Edition ed., Vol. 10, pp. 337–344). Oxford, UK: Elsevier.

Scobbie, J. M., Lawson, E., Nakai, S., Cleland, J., & Stuart-Smith, J. (2015). Onset vs. coda asymmetry in the articulation of English /r/. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK: The University of Glasgow. http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0704.pdf.

Scobbie, J. M., Punnoose, R., & Khattab, G. (2013). Articulating five liquids: A single speaker ultrasound study of Malayalam. In L. Spreafico & A. Vietti (Eds.), *Rhotics. New data and*

*perspectives* (pp. 99–124). Bozen-Bolzano, Italy: Bozen-Bolzano University Press.

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, *15*(3), 143–158. doi: 10.1250/ast.15.143

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73–80. doi: 10.3758/bf03206849

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320. doi: 10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*(3), 277–287. doi: 10.1016/S0168-0102(03)00214-1

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797–1805. doi: 10.1121/1.401660

Shaywitz, B. A., Shaywltz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., … Gore, J. C. (1995). Sex differences in the functional organization of the brain for language. *Nature*, *373*(6515), 607–609. doi: 10.1038/373607a0

Shriberg, L. D. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech, Language, and Hearing Research*, *36*(1), 105–140. doi: 10.1044/jshr.3601.105

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations*. https://arxiv.org/pdf/1409.1556.pdf.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*. http://CRAN.R-project.org/package=afex.

Smayda, K. E., Van Engen, K. J., Maddox, T. W., & Chandrasekaran, B. (2016). Audio-visual and meaningful semantic context enhancements in older and younger adults. *PLOS ONE*,

*11*(3), e0152773. doi: 10.1371/journal.pone.0152773

Smit, A. B. (1993). Phonologic error distributions in the Iowa-Nebraska articulation norms project: Consonant singletons. *Journal of Speech and Hearing Research*, *36*, 533–547. doi: 10.1044/jshr.3603.533

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, *55*(4), 779–798. doi: 10.1044/jshd.5504.779

Smith, B. J., Mielke, J., Magloughlin, L., & Wilbanks, E. (2019). Sound change and coarticulatory variability involving English /ɹ/. *Glossa: A Journal of General Linguistics*, *4(1)*(63), 1-51. doi: 10.5334/gjgl.650

Smith, C. L. (2007). Prosodic accomodation by French speakers to a non-native interlocutor. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16$^{th}$ International Congress of Phonetic Sciences* (pp. 1081–1084). Retrieved from http://icphs2007.de/conference/Papers/1116/1116.pdf

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. doi: 10.1097/00003446-200506000-00003

Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research*, *176*, 245–258. doi: 10.1016/S0079-6123(09)17615-X

Sproat, R., & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, *21*(3), 291–311. doi: 10.1016/S0095-4470(19)31340-3

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi: 10.3758/BF03207704

Steinberg, D. D., & Sciarini, N. V. (2013). *An introduction to pyscholinguistics* (2nd ed.). London, UK; New York, USA: Routledge.

Stent, A. J., Huffman, M. K., & Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, *50*(3), 163–178. doi: 10.1016/j.specom.2007.07.005

Stern, D., Spieker, S., Barnett, R., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, *10*(1), 1–15. doi: 10.1017/S0305000900005092

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1), 3–45. doi: 10.1016/S0095-4470(19)31520-7

Stevens, K. N. (1998). *Acoustic phonetics* (Vol. 30). Cambridge, MA, USA: MIT press.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, *19*(6-7), 455–501. doi: 10.1080/02699200500113558

Strycharczuk, P., Lloyd, S., & Scobbie, J. M. (2019, November 7-8). *Articulatory Variation in Southern British English Rhotics* [Conference abstract]. R-Atics 6 Colloqium, Paris, France. https://lpp.in2p3.fr/wp-content/uploads/Colloques/Brochure_r-atics6.pdf#page=49.

Stuart-Smith, J. (2007). A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16$^{th}$ International Congress of Phonetic Sciences* (pp. 1449–1452). http://www.icphs2007.de/conference/Papers/1307/1307.pdf.

Stuart-Smith, J., Lawson, E., & Scobbie, J. M. (2014). Derhoticisation in Scottish English: A sociophonetic journey. In C. Celata, S. Calmai, & P. Bertinetto (Eds.), *Advances in Sociophonetics* (pp. 59–96). Amsterdam: John Benjamins.

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1983). Audio-visual speech perception, lipreading, and artificial stimulation. In M. Lutman & M. Haggard (Eds.), *Hearing science and hearing disorders* (pp. 131–182). London, UK: Academic Press. doi: 10.1016/B978-0-12-460440-7.50010-7

Summerfield, Q., Bruce, V., Cowey, A., Ellis, A. W., & Perrett, D. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 71–78. doi: 10.1098/rstb.1992.0009

Sweet, H. (1877). *A handbook of phonetics* (Vol. 2). Oxford, UK: Clarendon Press.

Sweet, H. (1890). *A primer of phonetics*. Oxford, UK: Clarendon Press.

Tabain, M. (1998). Non-sibilant fricatives in English: Spectral information above 10 kHz.

*Phonetica*, *55*(3), 107–130. doi: 10.1159/000028427

Tang, L. Y. W., Hannah, B., Jongman, A., Sereno, J., Wang, Y., & Hamarneh, G. (2015). Examining visible articulatory features in clear and plain speech. *Speech Communication*, *75*, 1–13. doi: 10.1016/j.specom.2015.09.008

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850–855. doi: 10.1016/j.cognition.2008.05.009

Tiede, M. K., Boyce, S. E., Espy-Wilson, C. Y., & Gracco, V. L. (2010). Variability of North American English /r/ production in response to palatal perturbation. In B. Maassen & P. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research* (pp. 53–68). Oxford, UK: Oxford Scholarship Online. doi: 10.1093/acprof:oso/9780199235797.003.0004

Tiede, M. K., Boyce, S. E., Holland, C. K., & Choe, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *The Journal of the Acoustical Society of America*, *115*(5), 2633–2634. doi: 10.1121/1.4784878

Toda, M., Maeda, S., Carlen, A. J., & Meftahi, L. (2003). Lip protrusion/rounding dissociation in French and English consonants: /w/ vs. /ʃ/ and /ʒ/. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15$^{th}$ International Congress of Phonetic Sciences* (pp. 1763–1766). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1763.pdf.

Toutios, A., Lingala, S. G., Vaz, C., Kim, J., Esling, J., Keating, P., … Narayanan, S. S. (2016). Illustrating the production of the International Phonetic Alphabet sounds using fast real-time magnetic resonance imaging. *Proceedings of Interspeech 2016*, 2428–2432. doi: 10.21437/Interspeech.2016-605

Trask, R. L. (2004). *A dictionary of phonetics and phonology*. London, UK: Routledge.

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*(2), 244–258. doi: 10.1016/j.wocn.2006.03.002

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge University Press.

Trudgill, P. (1986). *Dialects in contact*. Oxford, UK: Blackwell.

Trudgill, P. (1988). Norwich revisited: Recent linguistic changes in an English urban dialect. *English World-Wide*, *9*, 33–49. doi: 10.1075/eww.9.1.03tru

Trudgill, P. (1999a). *The dialects of England* (2nd ed.). Oxford, UK: Blackwell.

Trudgill, P. (1999b). Norwich: Endogenous and exogenous linguistic change. In P. Foulkes & G. J. Docherty (Eds.), *Urban voices: Accent studies in the British Isles* (pp. 124–140). London, UK: Arnold.

Twist, A., Baker, A., Mielke, J., & Archangeli, D. (2007). Are "covert" /ɪ/ allophones really indistinguishable? *Penn Working Papers in Linguistics*, *13*(2). https://repository.upenn.edu/cgi/viewcontent.cgi?article=1014&context=pwpl.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, *18*(10), 883–892. doi: 10.3766/jaaa.18.10.7

Uldall, E. (1958). American 'molar' R and 'flapped' T. *Revista do Laboratório de Fonética Experimental da Faculdade de Letras da Universidade de Coimbra*, *4*, 103–106.

Underhill, A. (1994). *Sound foundations: Living phonology*. Oxford, UK: Heinemann.

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*(1), 2–7. doi: 10.1016/j.specom.2006.10.003

Vaissière, J. (2007). Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. In M. Solé, P. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 54–71). Oxford University Press.

Vaissière, J. (2009). Articulatory modeling and the definition of acoustic-perceptual targets for reference vowels. *The Chinese Phonetics Journal*(2), 22–33.

Vaissière, J. (2011). On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. *Proceedings of the 17$^{th}$ International Congress of Phonetic Sciences*, 52–59.

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of*

*Speech, Language, and Hearing Research*, *57*(5), 1908–1918. doi: 10.1044/JSLHR-H-13-0076

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396–403. doi: 10.3758/s13414-016-1238-9

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917–928. doi: 10.1121/1.396660

Vihman, M. (1996). *Phonological Development: The Origins of Language in the Child.* Oxford, UK: Blackwell.

Walden, B. E., Erdman, S. A., Montgomery, A. A., & Schwartz, D. M. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech, Language, and Hearing Research*, *24*(2), 207–216. doi: 10.1044/jshr.2402.207

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*(1), 130–145. doi: 10.1044/jshr.2001.130

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *The Journal of the Acoustical Society of America*, *100*(2), 1153–1162. doi: 10.1121/1.416300

Weenink, D. (2014). *Speech signal processing with Praat.* http://www.fon.hum.uva.nl/david/LOT/sspbook.pdf.

Weirich, M., & Fuchs, S. (2013). Palatal morphology can influence speaker-specific realizations of phonemic contrasts. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1894–1908. doi: 10.1044/1092-4388(2013/12-0217)

Wells, J. (1982). *Accents of English.* Cambridge University Press.

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *46*(4), 551–568. doi: 10.1037/h0084331

Werner, R. (2019, November 7-8). *An Experimental Investigation on Rhoticity and* /r/*-Sandhi*

*in Devon English* [Conference abstract]. R-Atics 6 Colloqium, Paris, France. https://lpp.in2p3.fr/wp-content/uploads/Colloques/Brochure_r-atics6.pdf#page=30.

Westbury, J. R., & Hashi, M. (1997). Lip-pellet positions during vowels and labial consonants. *Journal of Phonetics*, *25*(4), 405–419. doi: 10.1006/jpho.1997.0050

Westbury, J. R., Hashi, M., & Lindstrom, M. J. (1998). Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, *26*(3), 203–226. doi: 10.1016/S0167-6393(98)00058-2

Westbury, J. R., Turner, G., & Dembowski, J. (1994). *X-ray microbeam speech production database user's handbook.* Madison, WI, USA: University of Wisconsin. http://www.haskins.yale.edu/staff/gafos_downloads/ubdbman.pdf.

Widex. (n.d.). *Online hearing test.* https://www.widex.co.uk/en-gb/online-hearing-test/.

Williams, A., & Kerswill, P. (1999). Dialect levelling: Change and continuity in Milton Keynes, Reading and Hull. In P. Foulkes & G. J. Docherty (Eds.), *Urban voices: Accent studies in the British Isles* (pp. 141–162). London, UK: Arnold.

Wilson, A. H., Alsius, A., Paré, M., & Munhall, K. G. (2016). Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech, Language, and Hearing Research : JSLHR*, *59*(4), 601–615. doi: 10.1044/2016_JSLHR-S-15-0092

Wilson, I. L. (2006). *Articulatory Settings of French and English Monolingual and Bilingual Speakers* (PhD Thesis, University of British Columbia, Vancouver, Canada). https://open.library.ubc.ca/cIRcle/collections/ubctheses/831/items/1.0092890.

Winter, B. (2020). *Statistics for Linguists: An Introduction Using R.* London, UK: Taylor & Francis.

Wood, S. (1986). The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. *The Journal of the Acoustical Society of America*, *80*(2), 391–401. doi: 10.1121/1.394090

Woodward, M. F., & Barber, C. G. (1960). Phoneme perception in lipreading. *Journal of Speech & Hearing Research*, *3*, 212–222. doi: 10.1044/jshr.0303.212

Wrench, A. A., & Scobbie, J. M. (2003). Categorising vocalisation of English /l/ using EPG, EMA

and ultrasound. In S. Palethorpe & M. Tabain (Eds.), *Proceedings of the 6th International Seminar on Speech Production* (pp. 314–319). Sydney, Australia: Macquarie Centre for Cognitive Science. https://core.ac.uk/download/pdf/161925727.pdf.

Wrench, A. A., & Scobbie, J. M. (2016). Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016). *CASL Research Centre Working Paper*, WP-24, 1–14. https://eresearch.qmu.ac.uk/handle/20.500.12289/4367.

Zawadzki, P. A., & Kuehn, D. P. (1980). A cineradiographic study of static and dynamic aspects of American English /r/. *Phonetica*, *37*(4), 253–266. doi: 10.1159/000259995

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014. Lecture Notes in Computer Science* (pp. 818–833). Cham, Switzerland: Springer International Publishing. doi: 10.1007/978-3-319-10590-1_53

Zerling, J.-P. (1992). Frontal lip shape for French and English vowels. *Journal of Phonetics*, *20*(1), 3–14. doi: 10.1016/S0095-4470(19)30249-9

Zhang, Z., Boyce, S. E., Espy-Wilson, C. Y., & Tiede, M. K. (2003). Acoustic Strategies for Production of American English 'retroflex' /r/. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1125–1128). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1125.pdf.

Zhou, X., Espy-Wilson, C. Y., Boyce, S. E., Tiede, M. K., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, *123*(6), 4466–4481. doi: 10.1121/1.2902168

# Index

This is a partial index which lists the main occurrences of recurrent terms within the manuscript. Page numbers in **bold** correspond to the definition of an entry.

## H



## J



## L



## M



## N



## P



## Q



## R



## S

## T



## U



## V